ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

https://www.econometricsociety.org/

Econometrica, Vol. 88, No. 5 (September, 2020), 1859–1898

LEAVE-OUT ESTIMATION OF VARIANCE COMPONENTS

PATRICK KLINE
Department of Economics, UC Berkeley and NBER

RAFFAELE SAGGIO
Department of Economics, University of British Columbia and NBER

MIKKEL SØLVSTEN
Department of Economics, University of Wisconsin-Madison

Econometrica, Vol. 88, No. 5 (September, 2020), 1859–1898

LEAVE-OUT ESTIMATION OF VARIANCE COMPONENTS

PATRICK KLINE
Department of Economics, UC Berkeley and NBER

RAFFAELE SAGGIO
Department of Economics, University of British Columbia and NBER

MIKKEL SØLVSTEN
Department of Economics, University of Wisconsin-Madison

We propose leave-out estimators of quadratic forms designed for the study of linear models with unrestricted heteroscedasticity. Applications include analysis of variance and tests of linear restrictions in models with many regressors. An approximation algorithm is provided that enables accurate computation of the estimator in very large data sets. We study the large sample properties of our estimator allowing the number of regressors to grow in proportion to the number of observations. Consistency is established in a variety of settings where plug-in methods and estimators predicated on homoscedasticity exhibit first-order biases. For quadratic forms of increasing rank, the limiting distribution can be represented by a linear combination of normal and non-central χ^2 random variables, with normality ensuing under strong identification. Standard error estimators are proposed that enable tests of linear restrictions and the construction of uniformly valid confidence intervals for quadratic forms of interest. We find in Italian social security records that leave-out estimates of a variance decomposition in a two-way fixed effects model of wage determination yield substantially different conclusions regarding the relative contribution of workers, firms, and worker-firm sorting to wage inequality than conventional methods. Monte Carlo exercises corroborate the accuracy of our asymptotic approximations, with clear evidence of non-normality emerging when worker mobility between blocks of firms is limited.

KEYWORDS: Variance components, heteroscedasticity, fixed effects, leave-out estimation, many regressors, weak identification, random projection.

0. INTRODUCTION

AS ECONOMIC DATA SETS have grown large, so has the number of parameters employed in econometric models. Typically, researchers are interested in certain low dimensional summaries of these parameters that communicate the relative influence of the various economic phenomena under study. An important benchmark comes from Fisher's (1925) foundational work on analysis of variance (ANOVA) which he proposed as a means of achieving a "separation of the variance ascribable to one group of causes, from the variance ascribable to other groups."

Patrick Kline: pkline@econ.berkeley.edu
Raffaele Saggio: rsaggio@mail.ubc.ca
Mikkel Sølvsten: soelvsten@wisc.edu

We thank Isaiah Andrews, Bruce Hansen, Whitney Newey, Anna Mikusheva, Jack Porter, Andres Santos, Azeem Shaikh, and seminar participants at UC Berkeley, CEMFI, Chicago, Harvard, UCLA, MIT, Northwestern, NYU, Princeton, Queens, UC San Diego, Wisconsin, the NBER Labor Studies meetings, and the CEME Interactions workshop for helpful comments. The data used in this study were generously provided by the Fondazione Rodolfo De Benedetti and originally developed by the Economics Department of the Università Ca Foscari Venezia under the supervision of Giuseppe Tattara. We are grateful to the Berkeley Institute for Research on Labor and Employment for funding support and Schmidt Futures, which provided financial assistance for this project through the Labor Science Initiative at the Berkeley Opportunity Lab.

 https://doi.org/10.3982/ECTA16410

This paper develops a new approach to estimation of and inference on *variance components*, which we define broadly as quadratic forms in the parameters of a linear model. Traditional variance component estimators are predicated on the assumption that the errors in a linear model are identically distributed draws from a normal distribution. Standard references on this subject (e.g., Searle, Casella, and McCulloch (2009)) suggest diagnostics for heteroscedasticity and non-normality, but offer little guidance regarding estimation and inference when these problems are encountered. A closely related literature on panel data econometrics proposes variance component estimators designed for fixed effects models that restrict either the dimensionality of the underlying group means (Bonhomme, Lamadon, and Manresa (2019)) or the nature of the heteroscedasticity governing the errors (Andrews, Gill, Schank, and Upward (2008), Jochmans and Weidner (2019)).

Our first contribution is to propose a new variance component estimator designed for unrestricted linear models with heteroscedasticity of unknown form. The estimator is finite sample unbiased and can be written as a naive "plug-in" variance component estimator plus a bias correction term that involves "cross-fit" (Newey and Robins (2018)) estimators of observation-specific error variances. We also develop a representation of the estimator in terms of a covariance between outcomes and a "leave-one-out" generalized prediction (e.g., as in Powell, Stock, and Stoker (1989)). Building on work by Achlioptas (2003), we propose a random projection method that enables computation of our estimator in very large data sets with little loss of accuracy.

We study the asymptotic behavior of the proposed leave-out estimator in an environment where the number of regressors may be proportional to the sample size: a framework that has alternately been termed "many covariates" (Cattaneo, Jansson, and Newey (2018)) or "moderate dimensional" (Lei, Bickel, and El Karoui (2018)) asymptotics. Verifiable design requirements are provided that ensure the estimator is consistent. These design requirements are shown to be met in a series of examples where estimators relying on jackknife or homoscedasticity-based bias corrections are inconsistent.

Three sets of asymptotic results are developed that allow our estimator to be used for inference in a variety of settings. The first result concerns inference on quadratic forms of fixed rank, a problem that typically arises when testing a few linear restrictions in a model with many covariates (Cattaneo, Jansson, and Newey (2018)). Familiar examples include testing that particular parameters are significant in a fixed effects model and conducting inference on the coefficients from a projection of fixed effects onto a low dimensional vector of covariates. Extending classic proposals by Horn, Horn, and Duncan (1975) and MacKinnon and White (1985), we show that our leave-out approach can be used to construct an Eicker–White style variance estimator that is unbiased in the presence of unrestricted heteroscedasticity and that enables consistent inference on linear contrasts under weaker design restrictions than those considered by Cattaneo, Jansson, and Newey (2018).

Next, we derive a result establishing asymptotic normality for quadratic forms of growing rank. Such quadratic forms typically arise when conducting analysis of variance, but also feature in tests of model specification involving a large number of linear restrictions (Anatolyev (2012), Chao, Hausman, Newey, Swanson, and Woutersen (2014)). The large sample distribution of the estimator is derived using a variant of the arguments in Chatterjee (2008) and Sølvsten (2020). A consistent standard error estimator is proposed that utilizes sample splitting formulations of the sort considered by Newey and Robins (2018).

Finally, we present conditions under which the large sample distribution of our estimator is non-pivotal and can be represented by a linear combination of normal and non-central χ^2 random variables, with the non-centralities of the χ^2 terms serving as weakly identified nuisance parameters. This distribution arises in a two-way fixed effects model when there are "bottlenecks" in the mobility network. Such bottlenecks are shown to emerge, for example, when worker mobility is governed by a stochastic block model with limited mobility between blocks. To construct asymptotically valid confidence intervals in the presence of nuisance parameters, we propose inversion of a minimum distance test statistic. Critical values are obtained via an application of the procedure of Andrews and Mikusheva (2016). The resulting confidence interval is shown to be valid uniformly in the values of the nuisance parameters and to have a closed form representation in many settings, which greatly simplifies its computation.

We illustrate our results with an application of the two-way worker-firm fixed effects model of Abowd, Kramarz, and Margolis (1999) to Italian social security records. The proposed leave-out estimator finds a substantially smaller contribution of firms to wage inequality and a much stronger correlation between worker and firm effects than either the uncorrected plug-in estimator originally considered by Abowd, Kramarz, and Margolis (1999) or the homoscedasticity-based correction procedure of Andrews et al. (2008).

Projecting the estimated firm effects onto worker and firm characteristics, we find that older workers tend to be employed at firms offering higher firm wage effects and that this phenomenon is largely explained by the tendency of older workers to sort to bigger firms. Leave-out standard errors for the coefficients of these linear projections are found to be several times larger than a naive standard error predicated on the assumption that the estimated fixed effects are independent of each other. Stratifying our analysis by birth cohort, we formally reject the null hypothesis that older and younger workers face identical vectors of firm effects.

To assess the accuracy of our asymptotic approximations, we conduct a series of Monte Carlo exercises utilizing the realized mobility patterns of workers between firms. Clear evidence of non-normality arises in the sampling distribution of the estimated variance of firm effects in settings where the worker-firm mobility network is weakly connected. The proposed confidence intervals are shown to provide reliable size control in both strongly and weakly identified settings.

1. UNBIASED ESTIMATION OF VARIANCE COMPONENTS

Consider the linear model

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

where the regressors $x_i \in \mathbb{R}^k$ are non-random and the design matrix $S_{xx} = \sum_{i=1}^n x_i x_i'$ has full rank. The unobserved errors $\{\varepsilon_i\}_{i=1}^n$ are mutually independent and obey $\mathbb{E}[\varepsilon_i] = 0$, but may possess observation-specific variances $\mathbb{E}[\varepsilon_i^2] = \sigma_i^2$.

Our object of interest is a quadratic form $\theta = \beta' A \beta$ for some known non-random symmetric matrix $A \in \mathbb{R}^{k \times k}$ of rank r. Following Searle, Casella, and McCulloch (2009), when A is positive semi-definite, θ is a *variance* component, while when A is non-definite, θ may be referred to as a *covariance* component. Note that linear restrictions on the parameter vector β can be formulated in terms of variance components: for a non-random vector v, the null hypothesis $v'\beta = 0$ is equivalent to the restriction $\theta = 0$ when $A = vv'$. Examples from the economics literature where variance components are of direct interest are discussed in Section 2.

1.1. *Estimator*

A naive plug-in estimator of θ is given by the quadratic form $\hat{\theta}_{\text{PI}} = \hat{\beta}' A \hat{\beta}$, where $\hat{\beta} = S_{xx}^{-1} \sum_{i=1}^{n} x_i y_i = \beta + S_{xx}^{-1} \sum_{i=1}^{n} x_i \varepsilon_i$ denotes the ordinary least squares (OLS) estimator of β. Estimation error in $\hat{\beta}$ leads the plug-in estimator to exhibit a bias involving a linear combination of the unknown variances $\{\sigma_i^2\}_{i=1}^{n}$. Specifically, standard results on quadratic forms imply that

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

where $B_{ii} = x_i' S_{xx}^{-1} A S_{xx}^{-1} x_i$ measures the influence of the ith squared error ε_i^2 on $\hat{\theta}_{\text{PI}}$. As discussed in Section 2, this bias can be particularly severe when the dimension of the regressors k is large relative to the sample size.

To remove this bias, we develop leave-out estimators of the error variances $\{\sigma_i^2\}_{i=1}^{n}$. Denote the leave-i-out OLS estimator of β by $\hat{\beta}_{-i} = (S_{xx} - x_i x_i')^{-1} \sum_{\ell \neq i} x_\ell y_\ell$. An unbiased estimator of σ_i^2 is

$$\hat{\sigma}_i^2 = y_i\bigl(y_i - x_i' \hat{\beta}_{-i}\bigr).$$

We therefore propose the following bias-corrected estimator of θ:

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

While Newey and Robins (2018) observed that "cross-fit" covariances relying on sample splitting can be used to remove bias of the sort considered here, we are not aware of prior proposals to use the leave-i-out error variance estimators $\{\hat{\sigma}_i^2\}_{i=1}^{n}$.

One can also motivate $\hat{\theta}$ via a change of variables argument. Letting $\tilde{x}_i = A S_{xx}^{-1} x_i$ denote a vector of "generalized" regressors, we can write

$$\theta = \beta' A \beta = \beta' S_{xx} S_{xx}^{-1} A \beta = \sum_{i=1}^{n} \beta' x_i \tilde{x}_i' \beta = \sum_{i=1}^{n} \mathbb{E}\bigl[y_i \tilde{x}_i' \beta\bigr].$$

This observation suggests using the unbiased *leave-out* estimator

$$\hat{\theta} = \sum_{i=1}^{n} y_i \tilde{x}_i' \hat{\beta}_{-i}. \tag{2}$$

Note that direct computation of $\hat{\beta}_{-i}$ can be avoided by exploiting the representation

$$y_i - x_i' \hat{\beta}_{-i} = \frac{y_i - x_i' \hat{\beta}}{1 - P_{ii}},$$

where $P_{ii} = x_i' S_{xx}^{-1} x_i$ gives the leverage of observation i. Applying the Sherman–Morrison–Woodbury formula (Woodbury (1949), Sherman and Morrison (1950)), this representation also reveals that (1) and (2) are numerically equivalent:

$$y_i\tilde{x}_i'\hat{\beta}_{-i} = \underbrace{y_i\tilde{x}_i'S_{xx}^{-1}\sum_{\ell\neq i}x_\ell y_\ell}_{=y_i\tilde{x}_i'\hat{\beta}-B_{ii}y_i^2} + \underbrace{\frac{y_i\tilde{x}_i'S_{xx}^{-1}x_ix_i'S_{xx}^{-1}}{1-P_{ii}}\sum_{\ell\neq i}x_\ell y_\ell}_{=B_{ii}y_ix_i'\hat{\beta}_{-i}} = y_i\tilde{x}_i'\hat{\beta} - B_{ii}\hat{\sigma}_i^2.$$

A similar combination of a change of variables argument and a leave-one-out estimator was used by Powell, Stock, and Stoker (1989) in the context of weighted average derivatives. The JIVE estimators proposed by Phillips and Hale (1977) and Angrist, Imbens, and Krueger (1999) also use a leave-one-out estimator, though without the change of variables.[1]

REMARK 1: The $\{\hat{\sigma}_i^2\}_{i=1}^n$ can also be used to construct an unbiased variance estimator

$$\hat{\mathbb{V}}[\hat{\beta}] = S_{xx}^{-1}\left(\sum_{i=1}^n x_ix_i'\hat{\sigma}_i^2\right)S_{xx}^{-1}.$$

Though $\hat{\mathbb{V}}[\hat{\beta}]$ need not be positive semi-definite, Section 4 shows that it can be used to perform asymptotically valid inference on linear contrasts in settings where existing Eicker–White estimators fail. Specifically, using $\hat{\mathbb{V}}[\hat{\beta}]$ leads to valid inference under conditions where the estimators of Rao (1970) and Cattaneo, Jansson, and Newey (2018) do not exist (see, e.g., Horn, Horn, and Duncan (1975), Verdier (2017)).

REMARK 2: The quantity $\hat{\mathbb{V}}[\hat{\beta}]$ is closely related to the HC2 variance estimator of MacKinnon and White (1985). While the HC2 estimator employs observation-specific variance estimators $\hat{\sigma}_{i,\mathrm{HC2}}^2 = \frac{(y_i-x_i'\hat{\beta})^2}{1-P_{ii}}$, $\hat{\mathbb{V}}[\hat{\beta}]$ relies instead on $\hat{\sigma}_i^2 = \frac{y_i(y_i-x_i'\hat{\beta})}{1-P_{ii}}$.

REMARK 3: The leave-out estimator is easily adapted to settings where the data are organized into mutually exclusive and independent "clusters" within which the errors may be dependent (e.g., as in Moulton (1986)). The change of variables argument leading to (2) also implies that an estimator of the form $\sum_{i=1}^n y_i\tilde{x}_i'\hat{\beta}_{-c(i)}$ will be unbiased in such settings, where $\hat{\beta}_{-c(i)}$ is the OLS estimator obtained after leaving out all observations in the cluster to which observation i belongs. Appendix A provides an application.

1.2. *Relation to Existing Approaches*

As detailed in Section 2, several literatures make use of bias corrections nominally predicated on homoscedasticity. A common "homoscedasticity-only" estimator takes the form

$$\hat{\theta}_{\mathrm{HO}} = \hat{\beta}'A\hat{\beta} - \sum_{i=1}^n B_{ii}\hat{\sigma}_{\mathrm{HO}}^2, \tag{3}$$

[1]The object of interest in JIVE estimation is a *ratio* of quadratic forms $\beta_1'S_{xx}\beta_2/\beta_2'S_{xx}\beta_2$ in the two-equation model $y_{ij} = x_i'\beta_j + \varepsilon_{ij}$ for $j = 1, 2$. When no covariates are present, using leave-out estimators of both the numerator and denominator of this ratio yields the JIVE1 estimator of Angrist, Imbens, and Krueger (1999).

where $\hat{\sigma}^2_{\text{HO}} = \frac{1}{n-k}\sum_{i=1}^n (y_i - x_i'\hat{\beta})^2$ is the degrees-of-freedom corrected variance estimator. A sufficient condition for unbiasedness of $\hat{\theta}_{\text{HO}}$ is that there be no empirical covariance between σ_i^2 and (B_{ii}, P_{ii}). This restriction is in turn implied by the special cases of *homoscedasticity* where σ_i^2 does not vary with i or *balanced design* where (B_{ii}, P_{ii}) does not vary with i. In general, however, this estimator will be biased (see, e.g., Scheffe (1959), Chapter 10).

A second estimator, closely related to $\hat{\theta}$, relies upon a jackknife bias correction (Quenouille (1949)) of the plug-in estimator. This estimator can be written

$$\hat{\theta}_{\text{JK}} = n\hat{\theta}_{\text{PI}} - \frac{n-1}{n}\sum_{i=1}^n \hat{\theta}_{\text{PI},-i} \quad \text{where } \hat{\theta}_{\text{PI},-i} = \hat{\beta}_{-i}' A \hat{\beta}_{-i}.$$

We show in the Supplemental Material (Kline, Saggio, and Sølvsten (2020)) that the conventional jackknife can produce first-order biases in the opposite direction of the bias in the plug-in estimator. This problem is also shown to extend to recently proposed jackknife adaptations (Hahn and Newey (2004), Dhaene and Jochmans (2015)) designed for long panels.

1.3. *Finite Sample Properties*

We now study the finite sample properties of the leave-out estimator $\hat{\theta}$ and its infeasible analogue $\theta^* = \hat{\beta}' A \hat{\beta} - \sum_{i=1}^n B_{ii}\sigma_i^2$, which uses knowledge of the individual error variances. The following lemma establishes that $\hat{\theta}$ is unbiased whenever each of the leave-one-out estimators $\hat{\beta}_{-i}$ exists, which can equivalently be expressed as the requirement that $\max_i P_{ii} < 1$. This condition turns out to also be necessary for the existence of unbiased estimators, which highlights the need for additional restrictions on the model or sample whenever some leverages equal 1.

LEMMA 1: (i) *If* $\max_i P_{ii} < 1$, *then* $\mathbb{E}[\hat{\theta}] = \theta$.
(ii) *Unbiased estimators of* $\theta = \beta' A \beta$ *exist for all* A *if and only if* $\max_i P_{ii} < 1$.

See Appendix B for proofs.

Next, we establish that, when the errors are normal, the infeasible estimator θ^* is a weighted sum of a series of non-central χ^2 random variables. This second result provides a useful point of departure for our asymptotic approximations and highlights the important role played by the matrix

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

which encodes features of both the target parameter (as defined by A) and the design matrix S_{xx}.

Let $\lambda_1, \dots, \lambda_r$ denote the nonzero eigenvalues of $\tilde{A}$, where $\lambda_1^2 \geq \cdots \geq \lambda_r^2$ and each eigenvalue appears as many times as its algebraic multiplicity. We use Q to refer to the corresponding matrix of orthonormal eigenvectors so that $\tilde{A} = QDQ'$ where $D = \text{diag}(\lambda_1, \dots, \lambda_r)$. With these definitions,

$$\hat{\beta}' A \hat{\beta} = \sum_{\ell=1}^r \lambda_\ell \hat{b}_\ell^2,$$

where $\hat{b} = (\hat{b}_1, \dots, \hat{b}_r)' = Q' S_{xx}^{1/2} \hat{\beta}$ contains r linear combinations of the elements in $\hat{\beta}$. The random vector $\hat{b}$ and the eigenvalues $\lambda_1, \dots, \lambda_r$ are central to both the finite sample distribution provided below in Lemma 2 and the asymptotic properties of $\hat{\theta}$ as studied in Sections 4–6. Each eigenvalue of $\tilde{A}$ can be thought of as measuring how strongly θ depends on a particular linear combination of the elements in β relative to the difficulty of estimating that combination (as summarized by S_{xx}^{-1}).

LEMMA 2: *If* $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, *then*
(i) $\hat{b} \sim \mathcal{N}(b, \mathbb{V}[\hat{b}])$ *where* $b = Q' S_{xx}^{1/2} \beta$,
(ii) $\theta^* = \sum_{\ell=1}^{r} \lambda_\ell (\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell])$.

The distribution of θ^* is a sum of r potentially dependent non-central χ^2 random variables with non-centralities $b = (b_1, \dots, b_r)'$. In the special case of homoscedasticity ($\sigma_i^2 = \sigma^2$) and no signal ($b = 0$), we have that $\hat{b} \sim \mathcal{N}(0, \sigma^2 I_r)$, which implies that the distribution of θ^* is a weighted sum of r *independent* central χ^2 random variables. The weights are the eigenvalues of $\tilde{A}$, therefore consistency of θ^* follows whenever the sum of the squared eigenvalues converges to zero. The next subsection establishes that the leave-out estimator remains consistent when a signal is present ($b \neq 0$) and the errors exhibit unrestricted heteroscedasticity.

1.4. *Consistency*

We now drop the normality assumption and provide conditions under which $\hat{\theta}$ remains consistent. To accommodate high dimensionality of the regressors, we allow all parts of the model to change with n:

$$y_{i,n} = x_{i,n}' \beta_n + \varepsilon_{i,n} \quad (i = 1, \dots, n),$$

where $x_{i,n} \in \mathbb{R}^{k_n}$, $S_{xx,n} = \sum_{i=1}^{n} x_{i,n} x_{i,n}'$, $\mathbb{E}[\varepsilon_{i,n}] = 0$, $\mathbb{E}[\varepsilon_{i,n}^2] = \sigma_{i,n}^2$, and $\theta_n = \beta_n' A_n \beta_n$ for some sequence of known non-random symmetric matrices $A_n \in \mathbb{R}^{k_n \times k_n}$ of rank r_n. By treating $x_{i,n}$ and A_n as sequences of constants, all uncertainty derives from the disturbances $\{\varepsilon_{i,n} : 1 \leq i \leq n, n \geq 1\}$. This *conditional* perspective is common in the statistics literature on ANOVA (Scheffe (1959), Searle, Casella, and McCulloch (2009)) and allows us to be agnostic about the potential dependency among the $\{x_{i,n}\}_{i=1}^{n}$ and A_n.[2] Following standard practice, we drop the n subscript in what follows. All limits are taken as n goes to infinity unless otherwise noted.

ASSUMPTION 1: (i) $\max_i(\mathbb{E}[\varepsilon_i^4] + \sigma_i^{-2}) = O(1)$; (ii) *there exists a* $c < 1$ *such that* $\max_i P_{ii} \leq c$ *for all* n; *and* (iii) $\max_i (x_i' \beta)^2 = O(1)$.

Part (i) of this condition limits the thickness of the tails in the error distribution, as is typically required for OLS estimation (see, e.g., Cattaneo, Jansson, and Newey (2018), page 10). The bounds on $(x_i' \beta)^2$ and P_{ii} imply that $\hat{\sigma}_i^2$ has bounded variance. Part (iii) is a

[2] An *unconditional* analysis might additionally impose distributional assumptions on A_n and consider $\bar{\theta} = \beta' \mathbb{E}[A_n] \beta$ as the object of interest. The uncertainty in $\hat{\theta} - \bar{\theta}$ can always be decomposed into components attributable to $\hat{\theta} - \theta$ and $\theta - \bar{\theta}$. Because the behavior of $\theta - \bar{\theta}$ depends entirely on model choices, we leave such an analysis to future work.

technical condition that can be relaxed to allow $\max_i (x_i'\beta)^2$ to increase slowly with sample size as discussed further in Section 8. From (ii), it follows that $\frac{k}{n} \le c < 1$ for all n.

The following lemma establishes consistency of $\hat{\theta}$.

LEMMA 3: *If Assumption 1 and one of the following conditions hold, then* $\hat{\theta} - \theta \xrightarrow{p} 0$:

(i) *A is positive semi-definite*, $\theta = \beta' A\beta = O(1)$, *and* $\operatorname{trace}(\tilde{A}^2) = \sum_{\ell=1}^r \lambda_\ell^2 = o(1)$.

(ii) $A = \frac{1}{2}(A_1'A_2 + A_2'A_1)$, *where* $\theta_1 = \beta' A_1' A_1 \beta$ *and* $\theta_2 = \beta' A_2' A_2 \beta$ *satisfy* (*i*).

Part (i) of Lemma 3 establishes consistency of variance components given boundedness of θ and a joint condition on the design matrix S_{xx} and the matrix A.[3] Part (ii) shows that consistency of covariance components follows from consistency of variance components that dominate them via the Cauchy–Schwarz inequality, that is, $\theta^2 = (\beta' A_1' A_2 \beta)^2 \le \theta_1 \theta_2$. In several of the examples discussed in the next section, $\operatorname{trace}(\tilde{A}^2)$ is of order r/n^2, which is necessarily small in large samples. A more extensive discussion of primitive conditions that yield $\operatorname{trace}(\tilde{A}^2) = o(1)$ is provided in Section 8.

2. EXAMPLES

We now consider three commonly encountered empirical examples where our proposed estimation strategy provides an advantage over existing methods.

EXAMPLE 1—Analysis of Covariance: Since the work of Fisher (1925), it has been common to summarize the effects of experimentally assigned treatments on outcomes with estimates of variance components. Consider a data set composed of observations on N groups with T_g observations in the gth group. The "analysis of covariance" model posits that outcomes can be written

$$y_{gt} = \alpha_g + x_{gt}'\delta + \varepsilon_{gt} \quad (g = 1, \dots, N, t = 1, \dots, T_g \ge 2),$$

where α_g is a group effect and x_{gt} is a vector of strictly exogenous covariates.

A prominent example comes from Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) who studied the adult earnings y_{gt} of $n = \sum_{g=1}^N T_g$ students assigned experimentally to one of N different classrooms. Each student also has a vector of predetermined background characteristics x_{gt}. The variability in student outcomes attributable to classrooms can be written

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

where $\bar{\alpha} = \frac{1}{n}\sum_{g=1}^N T_g \alpha_g$ gives the (enrollment-weighted) mean classroom effect.

This model can be aligned with the notation of the preceding section by letting $i = i(g, t)$, where $i(\cdot, \cdot)$ is bijective, with inverse denoted $(g(\cdot), t(\cdot))$, and defining $y_i = y_{gt}$, $\varepsilon_i = \varepsilon_{gt}$,

$$x_i = (d_i', x_{gt}')', \qquad \beta = (\alpha', \delta')', \qquad \alpha = (\alpha_1, \dots, \alpha_N)' \quad \text{and} \quad d_i = (\mathbf{1}_{\{g=1\}}, \dots, \mathbf{1}_{\{g=N\}})'.$$

[3] A slight generalization of the proof reveals that this conclusion continues to hold under a locally misspecified model where $\max_i |\mathbb{E}[\varepsilon_i]| = O(1/\sqrt{n})$.

To represent the target parameter in this notation, we write $\sigma_\alpha^2 = \beta' A \beta$, where

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

Chetty et al. (2011) estimated σ_α^2 using a random effects ANOVA estimator (see, e.g., Searle, Casella, and McCulloch (2009)), which is of the homoscedasticity-only type given in (3). As shown in the Supplemental Material, this estimator is in general first-order biased when the errors are heteroscedastic and group sizes are unbalanced.

Special Case: No Common Regressors. When there are no common regressors ($x_{gt} = 0$ for all g, t), the leave-out estimator of σ_α^2 has a particularly simple representation:

$$\hat{\sigma}_\alpha^2 = \frac{1}{n}\sum_{g=1}^{N}\left(T_g(\hat{\alpha}_g - \hat{\bar{\alpha}})^2 - \left(1 - \frac{T_g}{n}\right)\hat{\sigma}_g^2\right) \quad \text{where } \hat{\sigma}_g^2 = \frac{1}{T_g - 1}\sum_{t=1}^{T_g}(y_{gt} - \hat{\alpha}_g)^2, \tag{4}$$

for $\hat{\alpha}_g = \frac{1}{T_g}\sum_{t=1}^{T_g} y_{gt}$ and $\hat{\bar{\alpha}} = \frac{1}{n}\sum_{g=1}^{N} T_g \hat{\alpha}_g$. This representation shows that if the model consists only of group-specific intercepts, then the leave-out estimator relies on group-level degrees-of-freedom corrections. The statistic in (4) was analyzed by Akritas and Papadatos (2004) in the context of testing the null hypothesis that $\sigma_\alpha^2 = 0$ while allowing for heteroscedasticity at the group level.

Covariance Representation. Another instructive representation of the leave-out estimator is in terms of the empirical covariance

$$\hat{\sigma}_\alpha^2 = \sum_{i=1}^{n} y_i \tilde{d}_i' \hat{\alpha}_{-i} \quad \text{where } \hat{\beta}_{-i} = \left(\hat{\alpha}_{-i}', \hat{\delta}_{-i}'\right).$$

The generalized regressor $\tilde{d}_i$ is a residual from an instrumental variables (IV) regression. Specifically, $\tilde{d}_i = \frac{1}{n}((d_i - \bar{d}) - \hat{\Gamma}'(x_{g(i)t(i)} - \bar{x}_{g(i)}))$, where $\bar{x}_g = \frac{1}{T_g}\sum_{t=1}^{T_g} x_{gt}$ and $\hat{\Gamma}$ gives the coefficients from an IV regression of $d_i - \bar{d}$ on $x_{g(i)t(i)} - \bar{x}_{g(i)}$ using $x_{g(i)t(i)}$ as an instrument. The IV residual $\tilde{d}_i$ is uncorrelated with $x_{g(i)t(i)}$ and has a covariance with d_i of $A_d' A_d$, which ensures that the empirical covariance between y_i and the generalized prediction $\tilde{d}_i{}' \hat{\alpha}_{-i}$ is an unbiased estimator of σ_α^2.

EXAMPLE 2—Random Coefficients: Group memberships are often modeled as influencing slopes in addition to intercepts (Kuh (1959), Hildreth and Houck (1968), Arellano and Bonhomme (2011)). Consider the following "random coefficient" model:

$$y_{gt} = \alpha_g + z_{gt}\gamma_g + \varepsilon_{gt} \quad (g = 1, \dots, N, t = 1, \dots, T_g \geq 3).$$

An influential example comes from Raudenbush and Bryk (1986), who modeled student mathematics scores as a "hierarchical" linear function of socioeconomic status (SES) with school-specific intercepts ($\alpha_g \in \mathbb{R}$) and slopes ($\gamma_g \in \mathbb{R}$). Letting $\bar{\gamma} = \frac{1}{n}\sum_{g=1}^{N} T_g \gamma_g$ for $n = \sum_{g=1}^{N} T_g$, the student-weighted variance of slopes can be written

$$\sigma_\gamma^2 = \frac{1}{n}\sum_{g=1}^{N} T_g(\gamma_g - \bar{\gamma})^2.$$

In the notation of the preceding section, we can write this model $y_i = x_i'\beta + \varepsilon_i$, where

$$x_i = (d_i', d_i' z_{gt})', \qquad \beta = (\alpha', \gamma')', \qquad \gamma = (\gamma_1, \dots, \gamma_N)' \quad \text{and} \quad \sigma_\gamma^2 = \beta' \begin{bmatrix} 0 & 0 \\ 0 & A_d' A_d \end{bmatrix} \beta$$

for y_i, ε_i, d_i, α, and A_d as defined in the preceding example.

Raudenbush and Bryk (1986) used a maximum likelihood estimator of σ_γ^2 predicated upon normality and homoscedastic errors. Swamy (1970) considered an estimator of σ_γ^2 that relies on group-level degrees-of-freedom corrections and is unbiased when the error variance is allowed to vary at the group level, but not with the level of z_{gt}. Arellano and Bonhomme (2011) proposed an estimator that is unbiased under arbitrary heteroscedasticity patterns, which by Lemma 1 implies the leverage requirement $\max_i P_{ii} < 1$. Our proposed leave-out estimator is also unbiased under arbitrary patterns of heteroscedasticity and takes a particularly simple form.

Covariance Representation. The leave-out estimator can be written

$$\hat{\sigma}_\gamma^2 = \sum_{i=1}^{n} y_i \tilde{z}_i \tilde{d}_i' \hat{\gamma}_{-i} \quad \text{where } \tilde{d}_i = \frac{1}{n}(d_i - \bar{d}), \tilde{z}_i = \frac{z_{g(i)t(i)} - \bar{z}_{g(i)}}{\sum_{t=1}^{T_{g(i)}} (z_{g(i)t} - \bar{z}_{g(i)})^2},$$

and $\bar{z}_g = \frac{1}{T_g} \sum_{t=1}^{T_g} z_{gt}$. Demeaning $z_{g(i)t(i)}$ at the group level guarantees $\tilde{d}_i \tilde{z}_i$ is uncorrelated with d_i, while scaling by the group variability in $z_{g(i)t}$ ensures that the covariance between $\tilde{d}_i \tilde{z}_i$ and $d_i z_{g(i)t(i)}$ is $A_d' A_d$.

EXAMPLE 3—Two-Way Fixed Effects: Economists often study settings where units possess two or more group memberships, some of which can change over time. A prominent example comes from Abowd, Kramarz, and Margolis (1999) (henceforth AKM) who proposed a panel model of log wage determination that is additive in worker and firm fixed effects. This so-called "two-way" fixed effects model takes the form

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + x_{gt}' \delta + \varepsilon_{gt} \quad (g = 1, \dots, N, t = 1, \dots, T_g \geq 2), \tag{5}$$

where the function $j(\cdot, \cdot) : \{1, \dots, N\} \times \{1, \dots, \max_g T_g\} \to \{0, \dots, J\}$ allocates each of $n = \sum_{g=1}^{N} T_g$ person-year observations to one of $J + 1$ firms. Here α_g is a "person effect," $\psi_{j(g,t)}$ is a "firm effect," x_{gt} is a time-varying covariate, and ε_{gt} is a time-varying error. In this context, the mean-zero assumption on the errors ε_{gt} can be thought of as requiring both the common covariates x_{gt} and the firm assignments $j(\cdot, \cdot)$ to obey a strict exogeneity condition.

Interest in such models often centers on understanding how much of the variability in log wages is attributable to firms. AKM summarized the firm contribution to wage inequality with the following two parameters:

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

where $\bar{\psi} = \frac{1}{n} \sum_{g=1}^{N} \sum_{t=1}^{T_g} \psi_{j(g,t)}$. The variance component σ_ψ^2 measures the direct contribution of firm wage variability to inequality, while the covariance component $\sigma_{\alpha,\psi}$ measures the additional contribution of systematic sorting of high wage workers to high wage firms.

To represent this model in the notation of the preceding section, define

$$x_i = (d_i', f_i', x_{gt}')', \qquad \beta = (\alpha', \psi', \delta')',$$
$$\alpha = (\alpha_1, \dots, \alpha_N)' + \mathbf{1}_N' \psi_0, \qquad \psi = (\psi_1 \dots, \psi_J)' - \mathbf{1}_J' \psi_0,$$

for y_i, ε_i, and d_i as in the preceding examples, and $f_i = (\mathbf{1}_{\{j(g,t)=1\}}, \dots, \mathbf{1}_{\{j(g,t)=J\}})'$. The target parameters are $\sigma_\psi^2 = \beta' A_\psi \beta$ and $\sigma_{\alpha,\psi} = \beta' A_{\alpha,\psi} \beta$, where

$$A_\psi = \begin{bmatrix} 0 & 0 & 0 \\ 0 & A_f' A_f & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A_{\alpha,\psi} = \frac{1}{2} \begin{bmatrix} 0 & A_d' A_f & 0 \\ A_f' A_d & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

for $A_f = \frac{1}{\sqrt{n}}(f_1 - \bar{f}, \dots, f_n - \bar{f})$, $\bar{f} = \frac{1}{n}\sum_{i=1}^n f_i$ and A_d as in the preceding examples.

Addition and subtraction of ψ_0 in β amounts to the normalization, $\psi_0 = 0$, which has no effect on the variance components of interest. As Abowd et al. (2002) noted, least squares estimation of (5) requires one normalization of the ψ vector within each set of firms connected by worker mobility. For simplicity, we assume all firms are connected so that only a single normalization is required.[4]

Covariance Representation. AKM estimated σ_ψ^2 and $\sigma_{\alpha,\psi}$ using the naive plug-in estimators $\hat{\beta}' A_\psi \hat{\beta}$ and $\hat{\beta}' A_{\alpha,\psi} \hat{\beta}$, which are biased. Andrews et al. (2008) proposed the "homoscedasticity-only" estimators of (3). These estimators are unbiased when the errors ε_i are independent and have common variance. Bonhomme, Lamadon, and Manresa (2019) proposed a two-step estimation approach that is consistent in the presence of heteroscedasticity when the support of firm wage effects is restricted to a finite number of values and each firm grows large with the total sample size n. Our leave-out estimators, which avoid both the homoscedasticity requirement on the errors and any cardinality restrictions on the support of the firm wage effects, can be written compactly as covariances taking the form

$$\hat{\sigma}_\psi^2 = \sum_{i=1}^n y_i x_i' S_{xx}^{-1} A_\psi \hat{\beta}_{-i}, \qquad \hat{\sigma}_{\alpha,\psi} = \sum_{i=1}^n y_i x_i' S_{xx}^{-1} A_{\alpha,\psi} \hat{\beta}_{-i}.$$

Notably, these estimators are unbiased whenever the leave-out estimator $\hat{\beta}_{-i}$ can be computed, regardless of the distribution of firm sizes.

Special Case: Two Time Periods. A simpler representation of $\hat{\sigma}_\psi^2$ is available in the case where only two time periods are available and no common regressors are present ($T_g = 2$ and $x_{gt} = 0$ for all g, t). Consider this model in first differences:

$$\Delta y_g = \Delta f_g' \psi + \Delta \varepsilon_g \quad (g = 1, \dots, N), \tag{6}$$

[4]Bonhomme, Lamadon, and Manresa (2019) studied a closely related model where workers and firms each belong to one of a finite number of types and each pairing of worker and firm type is allowed a different mean wage. These mean wage parameters are shown to be identified when each worker type moves between each firm type with positive probability, enabling estimation even when many firms are not connected.

where $\Delta y_g = y_{g2} - y_{g1}$, $\Delta\varepsilon_g = \varepsilon_{g2} - \varepsilon_{g1}$, and $\Delta f_g = f_{i(g,2)} - f_{i(g,1)}$. The leave-out estimator of σ^2_ψ applied to this differenced representation of the model is

$$\hat{\sigma}^2_\psi = \sum_{g=1}^{N} \Delta y_g \Delta\tilde{f}'_g \hat{\psi}_{-g} \quad \text{where } \Delta\tilde{f}_g = A_{ff} S^{-1}_{\Delta f \Delta f} \Delta f_g,$$

where the quantities $S_{\Delta f \Delta f}$ and $\hat{\psi}_{-g}$ correspond respectively to S_{xx} and $\hat{\beta}_{-i}$.

REMARK 4: The leave-out representation above reveals that $\hat{\sigma}^2_\psi$ is unbiased not only under arbitrary heteroscedasticity and design unbalance, but also under arbitrary correlation between ε_{g1} and ε_{g2}. The same can be shown to hold for $\hat{\sigma}_{\alpha,\psi}$. Furthermore, this representation highlights that $\hat{\sigma}^2_\psi$ only depends upon observations with $\Delta f_g \neq 0$ (i.e., firm "movers").

3. LARGE SCALE COMPUTATION

It is possible to quickly approximate $\hat{\theta}$ in large scale applications using a variant of the random projection method of Achlioptas (2003), which we refer to as the Johnson–Lindenstrauss Approximation (JLA) for its connection to the work of Johnson and Lindenstrauss (1984). JLA can be described by the following algorithm: fix a $p \in \mathbb{N}$ and generate the matrices $R_B, R_P \in \mathbb{R}^{p\times n}$, where (R_B, R_P) are composed of mutually independent Rademacher random variables that are independent of the data, that is, their entries take the values 1 and -1 with probability 1/2. Next, decompose A into $A = \frac{1}{2}(A'_1 A_2 + A'_2 A_1)$ for $A_1, A_2 \in \mathbb{R}^{n\times k}$, where $A_1 = A_2$ if A is positive semi-definite.[5] Let

$$\hat{P}_{ii} = \frac{1}{p}\left\| R_P X S^{-1}_{xx} x_i \right\|^2 \quad \text{and} \quad \hat{B}_{ii} = \frac{1}{p}\left(R_B A_1 S^{-1}_{xx} x_i\right)'\left(R_B A_2 S^{-1}_{xx} x_i\right),$$

where $X = (x_1, \ldots, x_n)'$. The Johnson–Lindenstrauss approximation to $\hat{\theta}$ is

$$\hat{\theta}_{JLA} = \hat{\beta}' A \hat{\beta} - \sum_{i=1}^{n} \hat{B}_{ii} \hat{\sigma}^2_{i,JLA},$$

where $\hat{\sigma}^2_{i,JLA} = \frac{y_i(y_i - x'_i\hat{\beta})}{1-\hat{P}_{ii}}\left(1 - \frac{1}{p}\frac{3\hat{P}^3_{ii}+\hat{P}^2_{ii}}{1-\hat{P}_{ii}}\right)$. The term $\frac{1}{p}\frac{3\hat{P}^3_{ii}+\hat{P}^2_{ii}}{1-\hat{P}_{ii}}$ removes a non-linearity bias introduced by approximating P_{ii}.[6]

The following lemma establishes asymptotic equivalence between the leave-out estimator $\hat{\theta}$ and its approximation $\hat{\theta}_{JLA}$ when p^4 is large relative to the sample size.

LEMMA 4: *If Assumption 1 is satisfied*, $n/p^4 = o(1)$, $\mathbb{V}[\hat{\theta}]^{-1} = O(n)$, *and one of the following conditions hold*, *then* $\mathbb{V}[\hat{\theta}]^{-1/2}(\hat{\theta}_{JLA} - \hat{\theta} - \mathrm{B}_p) = o_p(1)$ *where* $|\mathrm{B}_p| \leq \frac{1}{p}\sum_{i=1}^{n} P^2_{ii} \times |B_{ii}|\sigma^2_i$:

[5] In the examples of Section 2, A_1 and A_2 can be constructed using A_d and A_f.

[6] A MATLAB package (Kline, Saggio, and Sølvsten (2019b)) implementing both the exact and JLA versions of our estimator in the two-way fixed effects model is available online. The Computational Appendix, located in the replication file (Kline, Saggio, and Sølvsten (2020)), demonstrates that JLA allows us to accurately compute a variance decomposition in a two-way fixed effects model with roughly 15 million parameters—a scale comparable to the study of Card, Heining, and Kline (2013)—in under an hour.

(i) *A is positive semi-definite and* $\mathbb{E}[\hat{\beta}' A \hat{\beta}] - \theta = \sum_{i=1}^{n} B_{ii} \sigma_i^2 = O(1)$.
(ii) $A = \frac{1}{2}(A_1' A_2 + A_2' A_1)$ *where* $\theta_1 = \beta' A_1' A_1 \beta$ *and* $\theta_2 = \beta' A_2' A_2 \beta$ *satisfy* (i) *and* $\frac{\mathbb{V}[\hat{\theta}_1]\mathbb{V}[\hat{\theta}_2]}{n\mathbb{V}[\hat{\theta}]^2} = O(1)$.

For variance components, the lemma characterizes an approximation bias B_p in $\hat{\theta}_{JLA}$, which is at most $1/p$ times the bias in the plug-in estimator $\hat{\beta}' A \hat{\beta}$. For covariance components, asymptotic equivalence ensues when the variance components defined by $A_1' A_1$ and $A_2' A_2$ do not converge at substantially slower rates than $\hat{\theta}$. Under this condition, the approximation bias is at most $1/p$ times the average of the biases in the plug-in estimators $\hat{\beta}' A_1' A_1 \hat{\beta}$ and $\hat{\beta}' A_2' A_2 \hat{\beta}$.

These bounds on the approximation bias suggest that a p of a few hundred should suffice for point estimation. However, unless $n/p^2 = o(1)$, the resulting approximation bias needs to be accounted for when conducting inference. Specifically, one can lengthen the tails of the confidence sets proposed in Sections 5 and 7 by $\frac{1}{p} \sum_{i=1}^{n} \hat{P}_{ii}^2 |\hat{B}_{ii}| \hat{\sigma}_{i,JLA}^2$ when relying on JLA.

4. INFERENCE ON QUADRATIC FORMS OF FIXED RANK

While the examples of Section 2 emphasized variance components where the rank r of A was increasing with sample size, we first study the case where r is fixed. Problems of this nature often arise when testing a few linear restrictions or when conducting inference on linear combinations of the regression coefficients, say $v'\beta$. In the case of two-way fixed effects models of wage determination, the quantity $v'\beta$ might correspond to the difference in mean values of firm effects between male and female workers (Card, Cardoso, and Kline (2015)) or to the coefficient from a projection of firm effects onto firm size (Bloom, Guvenen, Smith, Song, and von Wachter (2018)). A third use case, discussed at length by Cattaneo, Jansson, and Newey (2018), is where $v'\beta$ corresponds to a linear combination of a few common coefficients in a linear model with high dimensional fixed effects that are regarded as nuisance parameters.

To characterize the limit distribution of $\hat{\theta}$ when r is small, we rely on a representation of θ as a weighted sum of squared linear combinations of the outcome data:

$$\hat{\theta} = \sum_{\ell=1}^{r} \lambda_\ell \big(\hat{b}_\ell^2 - \hat{\mathbb{V}}[\hat{b}_\ell]\big) \quad \text{where } \hat{b} = \sum_{i=1}^{n} w_i y_i \text{ and } \hat{\mathbb{V}}[\hat{b}] = \sum_{i=1}^{n} w_i w_i' \hat{\sigma}_i^2$$

for $w_i = (w_{i1}, \ldots, w_{ir})' = Q' S_{xx}^{-1/2} x_i$. The following theorem characterizes the asymptotic distribution of $\hat{\theta}$ while providing conditions under which $\hat{b}$ is asymptotically normal and $\hat{\mathbb{V}}[\hat{b}]$ is consistent.

THEOREM 1: *If Assumption 1 holds, r is fixed, and* $\max_i w_i' w_i = o(1)$, *then*
(i) $\mathbb{V}[\hat{b}]^{-1/2}(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, I_r)$ *where* $b = Q' S_{xx}^{1/2} \beta$,
(ii) $\mathbb{V}[\hat{b}]^{-1} \hat{\mathbb{V}}[\hat{b}] \xrightarrow{p} I_r$,
(iii) $\hat{\theta} = \sum_{\ell=1}^{r} \lambda_\ell (\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell]) + o_p(\mathbb{V}[\hat{\theta}]^{1/2})$.

The high-level requirement of this theorem that $\max_i w_i' w_i = o(1)$ is a Lindeberg condition ensuring that no observation is too influential. One can think of $\max_i w_i' w_i$ as measuring the inverse effective sample size available for estimating b: when the weights are equal across i, the equality $\sum_{i=1}^{n} w_i w_i' = I_r$ implies that $w_{i\ell}^2 = \frac{1}{n}$. Since $\frac{1}{n}\sum_{i=1}^{n} w_i' w_i = \frac{r}{n}$, the requirement that $\max_i w_i' w_i = o(1)$ is implied by a variety of primitive conditions that limit how far a maximum is from the average (see, e.g., Anatolyev (2012), Appendix A.1). Note that Theorem 1 does not apply to settings where r is proportional to n because $\max_i w_i' w_i \geq \frac{r}{n}$.

In the special case where $A = vv'$ for some non-random vector v, Theorem 1 establishes that the variance estimator $\hat{\mathbb{V}}[\hat{\beta}] = S_{xx}^{-1}(\sum_{i=1}^{n} x_i x_i' \hat{\sigma}_i^2) S_{xx}^{-1}$ enables consistent inference on the linear combination $v'\beta$ using the approximation

$$\frac{v'(\hat{\beta} - \beta)}{\sqrt{v'\hat{\mathbb{V}}[\hat{\beta}]v}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{7}$$

To derive this result we assumed that $\max_i P_{ii} \leq c$ for some $c < 1$, whereas standard Eicker–White variance estimators generally require that $\max_i P_{ii} \to 0$ and Cattaneo, Jansson, and Newey (2018) established an asymptotically valid approach to inference in settings where $\max_i P_{ii} \leq 1/2$. Thus, $\hat{\mathbb{V}}[\hat{\beta}]$ leads to valid inference under weaker conditions than existing versions of Eicker–White variance estimators.

REMARK 5: Theorem 1 extends classical results on hypothesis testing of a few linear restrictions, say, $H_0 : R\beta = 0$, to allow for many regressors and heteroscedasticity. A convenient choice of A for testing purposes is $\frac{1}{r}R'(RS_{xx}^{-1}R')^{-1}R$, where r, the rank of $R \in \mathbb{R}^{r\times k}$, is fixed. Under H_0, the asymptotic distribution of $\hat{\theta}$ is a weighted sum of r central χ^2 random variables. This distribution is known up to $\mathbb{V}[\hat{b}]$ and a critical value can be found through simulation.

5. INFERENCE ON QUADRATIC FORMS OF GROWING RANK

We now turn to the more challenging problem of conducting inference on θ when r increases with n, as in the examples discussed in Section 2. These results also enable tests of many linear restrictions. For example, in a model of gender-specific firm effects of the sort considered by Card, Cardoso, and Kline (2015), testing the hypothesis that men and women face identical sets of firm fixed effects entails as many equality restrictions as there are firms.

5.1. *Limit Distribution*

In order to describe the result, we introduce $\check{x}_i = \sum_{\ell=1}^{n} M_{i\ell} \frac{B_{\ell\ell}}{1-P_{\ell\ell}} x_\ell$, where $M_{i\ell} = \mathbf{1}_{\{i=\ell\}} - x_i S_{xx}^{-1} x_\ell$. Note that $\check{x}_i$ gives the residual from a regression of $\frac{B_{ii}}{1-P_{ii}} x_i$ on x_i. Therefore, $\check{x}_i = 0$ when the regressor design is balanced. The contribution of $\check{x}_i$ to the behavior of $\hat{\theta}$ is through the estimation of $\sum_{i=1}^{n} B_{ii}\sigma_i^2$, which can be ignored in the case where the rank of A is bounded. When the rank of A is large, as implied by condition (ii) of Theorem 2 below, this estimation error can resurface in the asymptotic distribution. One can think of the eigenvalue ratio in (ii) as the inverse effective rank of $\tilde{A}$: when all the eigenvalues are equal, $\frac{\lambda_1^2}{\sum_{\ell=1}^{r}\lambda_\ell^2} = \frac{1}{r}$.

THEOREM 2: *Recall that* $\tilde{x}_i = AS_{xx}^{-1}x_i$ *where* $\hat{\theta} = \sum_{i=1}^n y_i \tilde{x}_i' \hat{\beta}_{-i}$. *If Assumption* 1 *holds and the following conditions are satisfied*:

$$\text{(i)} \quad \mathbb{V}[\hat{\theta}]^{-1} \max_i \left(\left(\tilde{x}_i' \beta \right)^2 + \left(\check{x}_i' \beta \right)^2 \right) = o(1), \qquad \text{(ii)} \quad \frac{\lambda_1^2}{\sum_{\ell=1}^r \lambda_\ell^2} = o(1),$$

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

The proof of Theorem 2 relies on a variation of Stein's method developed in Sølvsten (2020) and a representation of $\hat{\theta}$ as a second order U-statistic, that is,

$$\hat{\theta} = \sum_{i=1}^n \sum_{\ell \neq i} C_{i\ell} y_i y_\ell, \tag{8}$$

where $C_{i\ell} = B_{i\ell} - 2^{-1} M_{i\ell}(M_{ii}^{-1} B_{ii} + M_{\ell\ell}^{-1} B_{\ell\ell})$ and $B_{i\ell} = x_i' S_{xx}^{-1} A S_{xx}^{-1} x_\ell$. The proof shows that the "kernel" $C_{i\ell}$ varies with n in such a way that $\hat{\theta}$ is asymptotically normal whether or not $\hat{\theta}$ is a degenerate U-statistic (i.e., whether or not β is zero).

One representation of the variance appearing in Theorem 2 is

$$\mathbb{V}[\hat{\theta}] = \sum_{i=1}^n \left(2\tilde{x}_i' \beta - \check{x}_i' \beta \right)^2 \sigma_i^2 + 2 \sum_{i=1}^n \sum_{\ell \neq i} C_{i\ell}^2 \sigma_i^2 \sigma_\ell^2.$$

Note that this variance is bounded from below by $\min_i \sigma_i^2 \sum_{i=1}^n (2\tilde{x}_i'\beta)^2 + (\check{x}_i'\beta)^2$ since $\sum_{i=1}^n \tilde{x}_i'\beta \check{x}_i'\beta = 0$. Therefore, condition (i) of Theorem 2 will be satisfied whenever $\max_i((\tilde{x}_i'\beta)^2 + (\check{x}_i'\beta)^2)$ is not too large compared to $\sum_{i=1}^n (\tilde{x}_i'\beta)^2 + (\check{x}_i'\beta)^2$. As in Theorem 1, (i) is implied by a variety of primitive conditions that limit how far a maximum is from the average, but since (i) involves a one-dimensional function of x_i it can also be satisfied when r is large. A particularly simple case where (i) is satisfied is when $\beta = 0$; further cases are discussed in Section 8.

REMARK 6: Theorem 2 can be used to test a large system of linear restrictions of the form $H_0 : R\beta = 0$, where $r \to \infty$ is the rank of $R \in \mathbb{R}^{r \times k}$. Under this null hypothesis, choosing $A = \frac{1}{r} R'(RS_{xx}^{-1}R')^{-1}R$ implies $\mathbb{V}[\hat{\theta}]^{-1/2}\hat{\theta} \xrightarrow{d} \mathcal{N}(0, 1)$ as all the nonzero eigenvalues of $\tilde{A}$ are equal to $\frac{1}{r}$. The existing literature allows for either heteroscedastic errors and moderately few regressors (Donald, Imbens, and Newey (2003), $k^3/n \to 0$) or homoscedastic errors and many regressors (Anatolyev (2012), $k/n \leq c < 1$). When coupled with the estimator of $\mathbb{V}[\hat{\theta}]$ presented in the next subsection, this result enables tests with heteroscedastic errors and many regressors.[7]

REMARK 7: Theorem 2 extends some common results in the literature on many and many weak instruments (see, e.g., Chao, Swanson, Hausman, Newey, and Woutersen

[7] This testing problem was also analyzed in Anatolyev and Sølvsten (2020), who proposed a finite sample correction to the critical value and a leave-three-out estimator of $\mathbb{V}[\hat{\theta}]$.

(2012)) where the estimators are asymptotically equivalent to quadratic forms. The structure of that setting is such that $\tilde{A} = I_r/r$ and $r \to \infty$, in which case condition (ii) of Theorem 2 is automatically satisfied.

5.2. *Variance Estimation*

In order to conduct inference based on the normal approximation in Theorem 2, we now propose an estimator of $\mathbb{V}[\hat{\theta}]$. The U-statistic representation of $\hat{\theta}$ in (8) implies that the variance of $\hat{\theta}$ is

$$\mathbb{V}[\hat{\theta}] = 4\sum_{i=1}^{n}\left(\sum_{\ell \neq i} C_{i\ell} x_\ell' \beta\right)^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{\ell \neq i} C_{i\ell}^2 \sigma_i^2 \sigma_\ell^2.$$

Naively replacing $\{x_i'\beta, \sigma_i^2\}_{i=1}^n$ with $\{y_i, \hat{\sigma}_i^2\}_{i=1}^n$ to form a plug-in estimator of $\mathbb{V}[\hat{\theta}]$ will, in general, lead to invalid inferences as $\hat{\sigma}_i^2\hat{\sigma}_\ell^2$ is a biased estimator of $\sigma_i^2\sigma_\ell^2$. For this reason, we consider estimators of the error variances that rely on leaving out more than one observation. We describe in the Supplemental Material a simple adjustment that leads to conservative inference in settings where leaving out more than one observation is infeasible.

Sample Splitting. To circumvent the aforementioned biases, we exploit two independent unbiased estimators of $x_i'\beta$ that are also independent of y_i. These estimators are denoted $\widehat{x_i'\beta}_{-i,s} = \sum_{\ell \neq i}^n P_{i\ell,s} y_\ell$ for $s = 1, 2$, where $P_{i\ell,s}$ does not (functionally) depend on the $\{y_i\}_{i=1}^n$. To ensure independence between $\widehat{x_i'\beta}_{-i,1}$ and $\widehat{x_i'\beta}_{-i,2}$, we require that $P_{i\ell,1}P_{i\ell,2} = 0$ for all ℓ. Employing these split sample estimators, we create a new set of unbiased estimators for σ_i^2:

$$\tilde{\sigma}_i^2 = \big(y_i - \widehat{x_i'\beta}_{-i,1}\big)\big(y_i - \widehat{x_i'\beta}_{-i,2}\big) \quad \text{and} \quad \hat{\sigma}_{i,-\ell}^2 = \begin{cases} y_i\big(y_i - \widehat{x_i'\beta}_{-i,1}\big) & \text{if } P_{i\ell,1} = 0, \\ y_i\big(y_i - \widehat{x_i'\beta}_{-i,2}\big) & \text{if } P_{i\ell,1} \neq 0, \end{cases}$$

where $\hat{\sigma}_{i,-\ell}^2$ is independent of y_ℓ and $\tilde{\sigma}_i^2$ is a cross-fit estimator of the form considered in Newey and Robins (2018). These cross-fit estimators can be used to construct an unbiased estimator of $\sigma_i^2\sigma_\ell^2$. Letting $P_{im,-\ell} = P_{im,1}1_{\{P_{i\ell,1}=0\}} + P_{im,2}1_{\{P_{i\ell,1}\neq 0\}}$ denote the weight observation m receives in $\hat{\sigma}_{i,-\ell}^2$ and $\tilde{C}_{i\ell} = C_{i\ell}^2 + 2\sum_{m=1}^n C_{mi}C_{m\ell}(P_{mi,1}P_{m\ell,2} + P_{mi,2}P_{m\ell,1})$, we define

$$\widehat{\sigma_i^2\sigma_\ell^2} = \begin{cases} \hat{\sigma}_{i,-\ell}^2 \cdot \hat{\sigma}_{\ell,-i}^2 & \text{if } P_{im,-\ell}P_{\ell m,-i} = 0 \text{ for all } m, \\ \tilde{\sigma}_i^2 \cdot \hat{\sigma}_{\ell,-i}^2 & \text{else if } P_{i\ell,1} + P_{i\ell,2} = 0, \\ \hat{\sigma}_{i,-\ell}^2 \cdot \tilde{\sigma}_\ell^2 & \text{else if } P_{\ell i,1} + P_{\ell i,2} = 0, \\ \hat{\sigma}_{i,-\ell}^2 \cdot (y_\ell - \bar{y})^2 \cdot 1_{\{\tilde{C}_{i\ell}<0\}} & \text{otherwise.} \end{cases}$$

The first three cases in the above definition correspond respectively to pairs where (i) $\hat{\sigma}_{i,-\ell}^2$ and $\hat{\sigma}_{\ell,-i}^2$ are independent, (ii) $\widehat{x_i'\beta}_{-i,1}$ and $\widehat{x_i'\beta}_{-i,2}$ are independent of y_ℓ, and (iii) $\widehat{x_\ell'\beta}_{-\ell,1}$ and $\widehat{x_\ell'\beta}_{-\ell,2}$ are independent of y_i. When any of these three cases apply, the resulting estimator of $\sigma_i^2\sigma_\ell^2$ will be unbiased. For the remaining set of pairs $\mathcal{B} = \{(i, \ell) : P_{im,-\ell}P_{\ell m,-i} \neq 0 \text{ for some } m, P_{i\ell,1} + P_{i\ell,2} \neq 0, P_{\ell i,1} + P_{\ell i,2} \neq 0\}$ that comprise the fourth case, we rely on an unconditional variance estimator which leads to a biased estimator of $\sigma_i^2\sigma_\ell^2$ and conservative inference.

Design Requirements. Constructing the above split sample estimators places additional requirements on the design matrix S_{xx}. We briefly discuss these requirements in the context of Examples 1, 2, and 3. In Example 1, leave-one-out estimation requires a minimum group size of two, whereas existence of $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$ requires groups sizes of at least three. Conservative inference can be avoided when the minimum group size is at least four. In Example 2, minimum group sizes of three and five are sufficient to ensure feasibility of leave-one-out estimation and existence of $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$, respectively. Conservativeness can be avoided with a minimum group size of seven.

In the first-differenced representation of Example 3, the predictions $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$ are associated with particular paths in the worker-firm mobility network and independence requires that these paths be edge-disjoint. Menger's theorem (Menger (1927)) implies that $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$ exists if the design matrix has full rank when any two observations are dropped. Menger's theorem also implies that conservativeness can be avoided if the design matrix has full rank when any three observations are dropped. In our application, we use Dijkstra's algorithm to find the paths that generate $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$.

Consistency. The following lemma shows that $\widehat{\sigma_i^2\sigma_\ell^2}$ can be utilized to construct an estimator of $\mathbb{V}[\hat{\theta}]$ that delivers consistent inference when sufficiently few pairs fall into $\mathcal{B}$ and provides conservative inference otherwise.

LEMMA 5: *For $s = 1, 2$, suppose that $\widehat{x_i'\beta}_{-i,s}$ satisfies (unbiasedness) $\sum_{\ell\neq i}^n P_{i\ell,s}x_\ell'\beta = x_i'\beta$, (sample splitting) $P_{i\ell,1}P_{i\ell,2} = 0$ for all ℓ, and (projection property) $\lambda_{\max}(P_sP_s') = O(1)$ where $P_s = (P_{i\ell,s})_{i,\ell}$ is the hat-matrix corresponding to $\widehat{x_i'\beta}_{-i,s}$. Let*

$$\hat{\mathbb{V}}[\hat{\theta}] = 4\sum_{i=1}^n\left(\sum_{\ell\neq i} C_{i\ell}y_\ell\right)^2\tilde{\sigma}_i^2 - 2\sum_{i=1}^n\sum_{\ell\neq i}\tilde{C}_{i\ell}\widehat{\sigma_i^2\sigma_\ell^2}.$$

(i) *If the conditions of Theorem 2 hold and $|\mathcal{B}| = O(1)$, then $\hat{\mathbb{V}}[\hat{\theta}]^{-1/2}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$.*
(ii) *If the conditions of Theorem 2 hold, then $\liminf_{n\to\infty}\mathbb{P}(\theta \in [\hat{\theta} \pm z_\alpha\hat{\mathbb{V}}[\hat{\theta}]^{1/2}]) \geq 1 - \alpha$, where z_α^2 denotes the $(1-\alpha)$th quantile of a central χ_1^2 random variable.*

The first term in $\hat{\mathbb{V}}[\hat{\theta}]$ is a plug-in estimator with expectation $\mathbb{V}[\hat{\theta}] + 2\sum_{i=1}^n\sum_{\ell\neq i}\tilde{C}_{i\ell}\sigma_i^2\sigma_\ell^2$. Hence, the second term is a bias correction that completely removes the bias when $\mathcal{B} = \emptyset$ and leaves a positive bias otherwise. The Supplemental Material establishes validity of an adjustment to $\hat{\mathbb{V}}[\hat{\theta}]$ that utilizes an upward biased unconditional variance estimator for observations where it is not possible to construct $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$.

REMARK 8: The purpose of the condition $|\mathcal{B}| = O(1)$ in the above lemma is to ensure that the bias of $\hat{\mathbb{V}}[\hat{\theta}]$ grows small with the sample size. Because the bias of $\hat{\mathbb{V}}[\hat{\theta}]$ is non-negative, inference based on $\hat{\mathbb{V}}[\hat{\theta}]$ remains valid even when this condition fails, as stated in the second part of Lemma 5. To gauge the conservatism of the standard error estimate, researchers may wish to calculate the fraction of pairs that belong to $\mathcal{B}$ along with the share of observations for which it is not possible to construct $\{\widehat{x_i'\beta}_{-i,s}\}_{s=1,2}$.

6. WEAKLY IDENTIFIED QUADRATIC FORMS OF GROWING RANK

When condition (ii) of Theorem 2 is violated, inference based on Lemma 5 can be misleading. For example, in two-way fixed effects models, it is possible that bottlenecks arise in the mobility network that lead the largest eigenvalues to dominate the others. Section 8 formalizes this idea in a stochastic block model where limited mobility between blocks generates bottlenecks.

This section provides a theorem that covers the case where some of the squared eigenvalues $\lambda_1^2, \dots, \lambda_r^2$ are large relative to their sum $\sum_{\ell=1}^r \lambda_\ell^2$. To interpret this assumption, recall that each eigenvalue of $\tilde{A}$ measures how strongly θ depends on a particular linear combination of the elements of β relative to the difficulty of estimating that combination (as summarized by S_{xx}^{-1}). From Lemma 3, $\text{trace}(\tilde{A}^2) = \sum_{\ell=1}^r \lambda_\ell^2$ governs the total variability in $\hat{\theta}$. Therefore, Theorem 3 covers the case where θ depends strongly on a few linear combinations of β that are imprecisely estimated relative to the overall sampling uncertainty in $\hat{\theta}$. The following assumption formalizes this setting.

ASSUMPTION 2: *There exist a* $c > 0$ *and a known and fixed* $q \in \{1, \dots, r-1\}$ *such that*

$$\frac{\lambda_{q+1}^2}{\sum_{\ell=1}^r \lambda_\ell^2} = o(1) \quad \textit{and} \quad \frac{\lambda_q^2}{\sum_{\ell=1}^r \lambda_\ell^2} \geq c \quad \textit{for all } n.$$

Assumption 2 defines q as the number of squared eigenvalues that are large relative to their sum. Equivalently, q indexes the number of nuisance parameters in b that are *weakly identified* relative to their influence on θ and the uncertainty in $\hat{\theta}$. In Section 7.2, we offer some guidance on choosing q in settings where it is unknown.

6.1. *Limit Distribution*

Given knowledge of q, we can split $\hat{\theta}$ into a known function of $\hat{\mathsf{b}}_q = (\hat{b}_1, \dots, \hat{b}_q)'$ and $\hat{\theta}_q$ where $\hat{b}_1, \dots, \hat{b}_q$ are OLS estimators of the weakly identified nuisance parameters:

$$\hat{\mathsf{b}}_q = \sum_{i=1}^n \mathsf{w}_{iq} y_i, \quad \mathsf{w}_{iq} = (w_{i1}, \dots, w_{iq})',$$

$$\hat{\theta}_q = \hat{\theta} - \sum_{\ell=1}^q \lambda_\ell \big(\hat{b}_\ell^2 - \hat{\mathbb{V}}[\hat{b}_\ell]\big), \quad \hat{\mathbb{V}}[\hat{b}] = \sum_{i=1}^n w_i w_i' \hat{\sigma}_i^2.$$

The main difficulty in proving the following theorem is to show that the joint distribution of $(\hat{\mathsf{b}}_q', \hat{\theta}_q)'$ is normal, which we do using the same variation of Stein's method that was employed for Theorem 2. The high-level conditions involve $\tilde{x}_{iq}$ and $\check{x}_{iq}$ which are the parts of $\tilde{x}_i$ and $\check{x}_i$ that pertain to $\hat{\theta}_q$ and are defined in the proof of Theorem 3.

THEOREM 3: *If* $\max_i \mathsf{w}_{iq}' \mathsf{w}_{iq} = o(1)$, $\mathbb{V}[\hat{\theta}_q]^{-1} \max_i((\tilde{x}_{iq}'\beta)^2 + (\check{x}_{iq}'\beta)^2) = o(1)$, *and Assumptions* 1 *and* 2 *hold*, *then*

(i) $\mathbb{V}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)']^{-1/2}((\hat{\mathsf{b}}_q', \hat{\theta}_q)' - \mathbb{E}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)']) \xrightarrow{d} \mathcal{N}(0, I_{q+1})$,

(ii) $\hat{\theta} = \sum_{\ell=1}^q \lambda_\ell (\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell]) + \hat{\theta}_q + o_p(\mathbb{V}[\hat{\theta}]^{1/2})$.

Theorem 3 provides an approximation to $\hat{\theta}$ in terms of a quadratic function of q asymptotically normal random variables and a linear function of one asymptotically normal random variable. Here, the non-centralities $\mathbb{E}[\hat{\mathsf{b}}_q] = (b_1, \dots, b_q)'$ serve as nuisance parameters that influence both θ and the shape of the limiting distribution of $\hat{\theta} - \theta$. The next section proposes an approach to dealing with these nuisance parameters that provides asymptotically valid inference on θ for any value of q.

6.2. *Variance Estimation*

In Theorem 3, the relevant variance is $\Sigma_q = \mathbb{V}[(\hat{\mathsf{b}}'_q, \hat{\theta}_q)']$, which can be written

$$\Sigma_q = \sum_{i=1}^n \begin{bmatrix} \mathsf{w}_{iq}\mathsf{w}'_{iq}\sigma_i^2 & 2\mathsf{w}_{iq}\left(\sum_{\ell \neq i} C_{i\ell q} x'_\ell \beta\right)\sigma_i^2 \\ 2\mathsf{w}'_{iq}\left(\sum_{\ell \neq i} C_{i\ell q} x'_\ell \beta\right)\sigma_i^2 & 4\left(\sum_{\ell \neq i} C_{i\ell q} x'_\ell \beta\right)^2 \sigma_i^2 + 2\sum_{\ell \neq i} C_{i\ell q}^2 \sigma_i^2 \sigma_\ell^2 \end{bmatrix},$$

where $C_{i\ell q}$ is defined in the proof of Theorem 3. Our estimator of this variance reuses the split sample estimators introduced for Theorem 2:

$$\hat{\Sigma}_q = \sum_{i=1}^n \begin{bmatrix} \mathsf{w}_{iq}\mathsf{w}'_{iq}\hat{\sigma}_i^2 & 2\mathsf{w}_{iq}\left(\sum_{\ell \neq i} C_{i\ell q} y_\ell\right)\tilde{\sigma}_i^2 \\ 2\mathsf{w}'_{iq}\left(\sum_{\ell \neq i} C_{i\ell q} y_\ell\right)\tilde{\sigma}_i^2 & 4\left(\sum_{\ell \neq i} C_{i\ell q} y_\ell\right)^2 \tilde{\sigma}_i^2 - 2\sum_{\ell \neq i} \tilde{C}_{i\ell q}^2 \widetilde{\sigma_i^2 \sigma_\ell^2} \end{bmatrix},$$

where $\tilde{C}_{i\ell q}$ and $\widetilde{\sigma_i^2 \sigma_\ell^2}$ are defined in the proof of the next lemma which shows consistency of this variance estimator.

LEMMA 6: *For $s = 1, 2$, suppose that $\widehat{x'_i\beta}_{-i,s}$ satisfies $\sum_{\ell \neq i}^n P_{i\ell,s} x'_\ell \beta = x'_i \beta$, $P_{i\ell,1} P_{i\ell,2} = 0$ for all ℓ, and $\lambda_{\max}(P_s P'_s) = O(1)$. If the conditions of Theorem 3 hold and $|\mathcal{B}| = O(1)$, then $\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}$.*

REMARK 9: As in the case of variance estimation for Theorem 2, it may be that the design does not allow for construction of the predictions $\widehat{x'_i\beta}_{-i,1}$ and $\widehat{x'_i\beta}_{-i,2}$ used in $\hat{\Sigma}_q$. The Supplemental Material describes an adjustment to $\hat{\Sigma}_q$ which has a positive definite bias and therefore leads to conservative inferences when coupled with the inference method discussed in the next section.

7. INFERENCE WITH NUISANCE PARAMETERS

We now develop a two-sided confidence interval for θ that delivers asymptotic size control conditional on a choice of q. Our proposal involves inverting a minimum distance statistic in $\hat{\mathsf{b}}_q$ and $\hat{\theta}_q$, which Theorem 3 implies are jointly normally distributed. To avoid the conservatism associated with standard projection methods (e.g., Dufour and Jasiak (2001)), we adjust the critical value downwards to deliver size control on θ rather than $\mathbb{E}[(\hat{\mathsf{b}}'_q, \hat{\theta}_q)']$. However, unlike in standard projection problems, θ is a nonlinear function of $\mathbb{E}[\hat{\mathsf{b}}_q]$. To accommodate this complication, we use a critical value proposed by Andrews and Mikusheva (2016) that depends on the curvature of the problem.

7.1. *Inference With Known q*

The confidence interval we consider is based on inversion of a minimum-distance statistic for $(\hat{\mathsf{b}}_q', \hat{\theta}_q)'$ using the critical value proposed in Andrews and Mikusheva (2016). For a specified level of confidence, $1-\alpha$, we consider the interval

$$\hat{C}^{\theta}_{\alpha,q} = \left[\min_{(\dot{b}_1,\dots,\dot{b}_q,\dot{\theta}_q)' \in \hat{\mathsf{E}}_{\alpha,q}} \sum_{\ell=1}^{q} \lambda_\ell \dot{b}_\ell^2 + \dot{\theta}_q, \max_{(\dot{b}_1,\dots,\dot{b}_q,\dot{\theta}_q)' \in \hat{\mathsf{E}}_{\alpha,q}} \sum_{\ell=1}^{q} \lambda_\ell \dot{b}_\ell^2 + \dot{\theta}_q \right],$$

where

$$\hat{\mathsf{E}}_{\alpha,q} = \left\{ \left(\mathsf{b}_q', \theta_q\right)' \in \mathbb{R}^{q+1} : \begin{pmatrix} \hat{\mathsf{b}}_q - \mathsf{b}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix}' \hat{\Sigma}_q^{-1} \begin{pmatrix} \hat{\mathsf{b}}_q - \mathsf{b}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix} \le z^2_{\alpha,\hat{\kappa}_q} \right\}.$$

The critical value function, $z_{\alpha,\kappa}$, depends on the maximal curvature, κ, of a certain manifold (exact definitions of $z_{\alpha,\kappa}$ and κ are given in the Supplemental Material). Heuristically, κ can be thought of as summarizing the influence of the nuisance parameter $\mathbb{E}[\hat{\mathsf{b}}_q]$ on the shape of $\hat{\theta}$'s limiting distribution. Accordingly, $z^2_{\alpha,0}$ is equal to the $(1-\alpha)$th quantile of a central χ^2_1 random variable. As $\kappa \to \infty$, $z^2_{\alpha,\kappa}$ approaches the $(1-\alpha)$th quantile of a central χ^2_{q+1} random variable. This upper limit on $z_{\alpha,\kappa}$ is used in the projection method in its classical form as popularized in econometrics by Dufour and Jasiak (2001), while the lower limit z_α would yield size control if θ were linear in $\mathbb{E}[\hat{\mathsf{b}}_q]$.

When $q=1$, the maximal curvature is $\hat{\kappa}_1 = \frac{2|\lambda_1|\hat{\mathbb{V}}[\hat{b}_1]}{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}(1-\hat{\rho}^2)^{1/2}}$, where $\hat{\rho}$ is the estimated correlation between $\hat{b}_1$ and $\hat{\theta}_1$. This curvature measure is intimately related to eigenvalue ratios previously introduced, as $\hat{\kappa}_1^2$ is approximately equal to $\frac{2\lambda_1^2}{\sum_{\ell=2}^{r}\lambda_\ell^2}$ when the error terms are homoscedastic and $\beta = 0$. A closed form expression for the $q=1$ confidence interval is provided in the Supplemental Material. When $q > 1$, inference relies on solving two quadratic optimization problems that involve $q+1$ unknowns, which can be achieved reliably using standard quadratic programming routines.

The following lemma shows that a consistent variance estimator as proposed in Lemma 6 suffices for asymptotic validity under the conditions of Theorem 3.

LEMMA 7: *If $\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}$ and the conditions of Theorem 3 hold, then*

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta \in \hat{C}^{\theta}_{\alpha,q}\right) \ge 1-\alpha.$$

The confidence interval studied in Lemma 7 constructs a $q+1$ dimensional ellipsoid $\hat{\mathsf{E}}_{\alpha,q}$ and maps it through the quadratic function $(\dot{b}_1,\dots,\dot{b}_q,\dot{\theta}_q) \mapsto \sum_{\ell=1}^{q}\lambda_\ell \dot{b}_\ell^2 + \dot{\theta}_q$. This approach ensures uniform coverage over any possible values of the nuisance parameters $b_1,\dots,b_q$.

7.2. *Choosing q*

The preceding discussion of inference considered a setting where the number of weakly identified parameters was known in advance. When it is not clear what value q takes, researchers may wish to report confidence intervals for two consecutive values of q (or their union).[8] This observation also suggests a heuristic rule for choosing q: select q so that $\lambda_q^2/\sum_{\ell=1}^r \lambda_\ell^2 \geq \frac{1}{10}$ and $\lambda_{q+1}^2/\sum_{\ell=1}^r \lambda_\ell^2 < \frac{1}{10}$, with $q=0$ when $\lambda_1^2/\sum_{\ell=1}^r \lambda_\ell^2 < \frac{1}{10}$. A similar threshold rule can be motivated by a mild strengthening of Assumption 2 that allows one to learn q from the data.

ASSUMPTION 2′: *There exist a $c>0$, an $\epsilon>0$, and a fixed $q \in \{1,\dots,r-1\}$ such that*

$$\frac{\lambda_{q+1}^2}{\sum_{\ell=1}^r \lambda_\ell^2} = O\left(r^{-\varepsilon}\right) \quad \text{and} \quad \frac{\lambda_q^2}{\sum_{\ell=1}^r \lambda_\ell^2} \geq c \quad \text{for all } n.$$

A threshold based choice of q is the unique $\hat{q}$ for which

$$\frac{\lambda_{\hat{q}+1}^2}{\sum_{\ell=1}^r \lambda_\ell^2} < c_r \quad \text{and} \quad \frac{\lambda_{\hat{q}}^2}{\sum_{\ell=1}^r \lambda_\ell^2} \geq c_r \quad \text{for some } c_r \to 0,$$

with $\hat{q}=0$ when $\frac{\lambda_1^2}{\sum_{\ell=1}^r \lambda_\ell^2} < c_r$. Under Assumption 2′, $\hat{q}=q$ in sufficiently large samples provided that c_r is chosen so that $c_r r^\varepsilon \to \infty$. This condition is satisfied when c_r shrinks slowly to zero, for example, when $c_r \propto 1/\log(r)$.

8. VERIFYING CONDITIONS

We now revisit the examples of Section 2 and verify the conditions required to apply our theoretical results. The Supplemental Material provides further details on these calculations.

EXAMPLE 1—Analysis of Covariance, continued: Recall that $\theta = \sigma_\alpha^2 = \frac{1}{n}\sum_{g=1}^N T_g(\alpha_g - \bar{\alpha})^2$, where $y_{gt} = \alpha_g + x_{gt}'\delta + \varepsilon_{gt}$, g index the N groups, and T_g is group size.

No Common Regressors. Assumption 1(ii), (iii) requires $\max_g \alpha_g^2 = O(1)$ and $T_g \geq 2$ since $P_{ii} = T_{g(i)}^{-1}$. Consistency follows from Lemma 3 since $\lambda_\ell = \frac{1}{n}$ for $\ell = 1,\dots,r$ where $r = N-1$. Thus, $\operatorname{trace}(\tilde{A}^2) = r/n^2 \leq 1/n = o(1)$. Theorem 1 applies if the number of groups is fixed and $\min_g T_g \to \infty$, while Theorem 2 applies if the number of groups is large. Theorem 3 cannot apply as all eigenvalues are equal to $\frac{1}{n}$.

Common Regressors. To accommodate common regressors of fixed dimension, assume $\|\delta\|^2 + \max_{g,t}\|x_{gt}\|^2 = O(1)$ and that $\frac{1}{n}\sum_{g=1}^N \sum_{t=1}^{T_g}(x_{gt} - \bar{x}_g)(x_{gt} - \bar{x}_g)'$ converges to a positive definite limit. This is a standard assumption in basic panel data models (see, e.g., Wooldridge (2010), Chapter 10). Allowing such common regressors does not alter the previous conclusions: Theorem 1 applies if N is fixed and $\min_g T_g \to \infty$ since $w_i'w_i \leq P_{ii} = T_{g(i)}^{-1} + O(n^{-1})$, Theorem 2 applies if $N \to \infty$ since $\sum_{\ell=1}^n |M_{i\ell}| = O(1)$ implies that $\mathbb{V}[\hat{\theta}]^{-1}\max_i(\check{x}_i'\beta)^2 = o(1)$, and Theorem 3 cannot apply since $n\lambda_\ell \in [c_1, c_2]$ for $\ell = 1,\dots,r$ and some $c_2 \geq c_1 > 0$ not depending on n.

Unbounded Mean Function. All conclusions continue to hold if $\max_{g,t}\alpha_g^2 + \|x_{gt}\|^2 = O(1)$ is replaced with $\frac{\max_{g,t}\alpha_g^2 + \|x_{gt}\|^2}{\max\{N, \min_g T_g\}} = o(1)$ and $\sigma_\alpha^2 + \frac{1}{n}\sum_{g=1}^N \sum_{t=1}^{T_g}\|x_{gt}\|^2 = O(1)$.

[8]Both our simulations and empirical application suggest that $\hat{C}_{\alpha,q}^\theta$ barely varies with q when $\lambda_{q+1}^2/\sum_{\ell=1}^r \lambda_\ell^2 < \frac{1}{10}$. Consequently, little power is sacrificed by taking the union.

EXAMPLE 2—Random Coefficients, continued: For simplicity, consider the *uncentered* second moment $\theta = \frac{1}{n}\sum_{g=1}^{N} T_g \gamma_g^2$ where $y_{gt} = \alpha_g + z_{gt}'\gamma_g + \varepsilon_{gt}$. Suppose Assumption 1 holds and assume that $\max_{g,t} \alpha_g + \gamma_g^2 + z_{gt}^2 = O(1)$ and $\min_g S_{zz,g} \geq c > 0$, where $S_{zz,g} = \sum_{t=1}^{T_g}(z_{gt} - \bar{z}_g)^2$. Note that $\min_g S_{zz,g} > 0$ is equivalent to full rank of S_{xx} and $S_{zz,g}$ indexes how precisely γ_g can be estimated.

Consistency. The N eigenvalues of $\tilde{A}$ are $\lambda_g = \frac{T_g}{n} S_{zz,g}^{-1}$ for $g = 1, \dots, N$, where the group indexes are ordered so that $\lambda_1 \geq \cdots \geq \lambda_N$. Consistency follows from Lemma 3 if $\lambda_1^{-1} = n\frac{S_{zz,1}}{T_1} \to \infty$. This is automatically satisfied with many groups of bounded size.

Limit Distribution. If N is fixed and $\min_g S_{zz,g} \to \infty$, then Theorem 1 applies. If $\frac{\sqrt{N}}{T_1} S_{zz,1} \to \infty$, then Theorem 2 applies. If $\frac{\sqrt{N}}{T_2} S_{zz,2} \to \infty$, $\frac{\sqrt{N}}{T_1} S_{zz,1} = O(1)$, and $S_{zz,1} \to \infty$, then Theorem 3 applies with $q = 1$. In this case, γ_1 is weakly identified relative to its influence on θ and the overall variability of $\hat{\theta}$. This is expressed through the condition $\frac{\sqrt{N}}{T_1} S_{zz,1} = O(1)$, where $S_{zz,1}$ is the identification strength of γ_1, T_1 is the influence of γ_1 on θ, and $1/\sqrt{N}$ indexes the variability of $\hat{\theta}$.

EXAMPLE 3—Two-Way Fixed Effects, continued: In this final example, we restrict attention to the first-differenced setting $\Delta y_g = \Delta f_g'\psi + \Delta\varepsilon_g$ with $T_g = 2$ and a large number of firms, $J \to \infty$. Our target parameter is the variance of firm effects $\theta = \sigma_\psi^2 = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}(\psi_{j(g,t)} - \bar{\psi})^2$ and we consider Assumption 1 satisfied; in particular, $\max_j |\psi_j| = O(1)$.

Leverages. The leverage P_{gg} of observation g is less than 1 if the origin and destination firms of worker g are connected by a path not involving g. Letting n_g denote the number of edges in the shortest such path, one can show that $P_{gg} \leq \frac{n_g}{1+n_g}$. Therefore, if $\max_g n_g < 100$, then Assumption 1(ii) is satisfied with $\max P_{gg} \leq 0.99$. In our application, we find $\max_g n_g = 12$, leading to a somewhat smaller bound on the maximal leverage. The same consideration implies a bound on the model in levels since $P_{i(g,t)i(g,t)} = \frac{1}{2}(1 + P_{gg})$.

Eigenvalues. The eigenvalues of $\tilde{A}$ satisfy the equality

$$\lambda_\ell = \frac{1}{n\dot{\lambda}_{J+1-\ell}} \quad \text{for } \ell = 1, \dots, J,$$

where $\dot{\lambda}_1 \geq \cdots \geq \dot{\lambda}_J$ are the nonzero eigenvalues of the matrix $E^{1/2}\mathcal{L}E^{1/2}$. $\mathcal{L}$ is the normalized Laplacian of the employer mobility network and connectedness of the network is equivalent to full rank of S_{xx} (see the Supplemental Material for definitions). E is a diagonal matrix of employer-specific "churn rates," that is, the number of moves in and out of a firm divided by the total number of employees in the firm. E and $\mathcal{L}$ interact in determining the eigenvalues of $\tilde{A}$. In Example 2, the quantities $\{T_\ell^{-1} S_{zz,\ell}\}_{\ell=1}^{N}$ played a role directly analogous to the churn rates in E, so in this example we focus on the role of $\mathcal{L}$ by assuming that the diagonal entries of E are all equal to 1.

Strongly Connected Network. The employer mobility network is *strongly connected* if $\sqrt{J}\mathcal{C} \to \infty$, where $\mathcal{C} \in (0, 1]$ is Cheeger's constant for the mobility network (see, e.g., Mohar (1989), Jochmans and Weidner (2019)). Intuitively, $\mathcal{C}$ measures the most severe "bottleneck" in the network, where a bottleneck is a set of movers that, upon removal from the data, splits the mobility network into two disjoint blocks. The severity of the bottleneck is governed by the number of movers removed divided by the smallest number of movers in either of the two disjoint blocks. The inequalities $\dot{\lambda}_J \geq 1 - \sqrt{1-\mathcal{C}^2}$ (Chung (1997), Theorem 2.3) and $\lambda_1^2/\sum_{\ell=1}^J \lambda_\ell^2 \leq 4(\sqrt{J}\dot{\lambda}_J)^{-2}$ imply that a strongly connected network yields $q = 0$, which rules out application of Theorem 3. Furthermore, a strongly connected network is sufficient (but not necessary) for consistency of $\hat{\theta}$ as $\sum_{\ell=1}^J \lambda_\ell^2 \leq \frac{J}{n}(\sqrt{n}\dot{\lambda}_J)^{-2}$.

Weakly Connected Network. When $\sqrt{J}\mathcal{C}$ is bounded, the network is *weakly connected* and can contain a sufficiently severe bottleneck that a linear combination of the elements of ψ is estimated imprecisely relative to its influence on θ and the total uncertainty in $\hat{\theta}$. The weakly identified linear combination in this case is a difference in average firm effects across the two blocks on either side of the bottleneck, which contributes a χ^2 term to the asymptotic distribution. Below, we use a stochastic block model to further illustrate this phenomenon. Our empirical application demonstrates that weakly connected networks can appear in practically relevant settings.

Stochastic Block Model. Consider a stochastic block model of network formation where firms belong to one of two blocks and a set of workers switch firms, possibly by moving between blocks. Workers' mobility decisions are independent: with probability p_b a worker moves between blocks and with probability $1 - p_b$ she moves within block. For simplicity, we further assume that the two blocks contain equally many firms. To ensure Assumption 1(ii) holds, we work with a semi-sparse network where $\frac{J\log(J)}{n} + \frac{\log(J)}{np_b} \to 0$.[9] In this model, the asymptotic behavior of $\hat{\theta}$ is governed by p_b: the most severe bottleneck is between the two blocks and has a Cheeger's constant proportional to p_b. In the Supplemental Material, we use this model to verify the high-level conditions leading to Theorems 2 and 3 and show that Theorem 2 applies when $\sqrt{J}p_b \to \infty$, while Theorem 3 applies with $q = 1$ otherwise. The argument extends to any finite number of blocks, in which case q is the number of blocks minus one. Finally, we show that $\hat{\theta}$ is consistent even when the network is weakly connected. To establish consistency, we only impose $\frac{\log(J)}{np_b} \to 0$, which requires that the number of movers across the two blocks is large.

9. APPLICATION

In this section, we use Italian social security records to compute leave-out estimates of the AKM wage decomposition and contrast them with estimates based upon the plug-in estimator of Abowd, Kramarz, and Margolis (1999) and the homoscedasticity-corrected estimator of Andrews et al. (2008). We then investigate whether the variance components that comprise the AKM decomposition differ across age groups and test some hypotheses regarding the relationship between firm fixed effects, worker age, and firm size. We conclude with a Monte Carlo analysis of the performance of our proposed confidence intervals.

9.1. *Sample Construction*

The data used in our analysis come from the Veneto Worker History (VWH) file, which provides the annual earnings and days worked associated with each covered employment spell taking place in the Veneto region of Northeast Italy over the years 1984–2001. Our baseline sample consists of workers with employment spells taking place in the years 1999

[9]The semi-sparse stochastic block model is routinely employed in the literature on spectral clustering (e.g., Sarkar and Bickel (2015)) to guarantee connectedness of the network. The bias of the plug-in estimator under this model is $o(1/\log(J))$. Hence, bias correction is essential for valid inference but not for consistency.

TABLE I

SUMMARY STATISTICS[a]

	Pooled	Younger Workers	Older Workers
Largest Connected Set			
Number of Observations	1,859,459	1,027,034	643,020
Number of Movers	197,572	133,627	53,035
Number of Firms	73,933	62,848	26,606
Mean Log Daily Wage	4.7507	4.6694	4.8925
Variance Log Daily Wage	0.1985	0.1329	0.2722
Leave-one-out Sample			
Number of Observations	1,319,972	661,528	425,208
Number of Movers	164,203	102,746	35,467
Number of Firms	42,489	33,151	10,733
Mean Log Daily Wage	4.8066	4.7275	4.9455
Variance Log Daily Wage	0.1843	0.1200	0.2591
Maximum Leverage (P_{ii})	0.9365	0.9437	0.9513

[a]Data in column (1) correspond to VHW observations in the years 1999 and 2001. Column (2) restricts to workers born in the years 1965–1983. Column (3) considers workers born in the years 1937–1964. The largest connected set gives the largest sample in which all firms are connected by worker mobility. The leave-one-out sample is the largest connected set such that every firm remains connected after removing any single worker from the sample. Statistics on log daily wages are person-year weighted.

and 2001, which provides us with a three-year horizon over which to measure job mobility. A longer panel, spanning the years 1996–2001, is studied in Appendix A. For each worker, we retain the employment spells yielding the highest earnings in the years 1999 and 2001. Wages in each year are defined as earnings in the selected spell divided by the spell length in days. Workers are divided into two groups of roughly equal size according to their year of birth: "younger" workers born in the years 1965–1983 (aged 18–34 in 1999) and "older" workers born in the years 1937–1964 (aged 35–64 in 1999). Further details on our processing of the VWH records are provided in the Computational Appendix.

Table I reports the number of person-year observations available among workers employed by firms in the region's largest connected set, along with the largest connected set for each age group. Workers are classified as "movers" if they switch firms between 1999 and 2001. Roughly 21% of all workers are movers and the average number of movers per connected firm ranges from nearly 3 in the pooled sample to roughly 2 in the thinner age-specific samples.

Our leave-out estimation strategy requires that firms remain connected by worker mobility when any single mover is dropped. Pruning the sample to ensure this condition holds drops roughly half of the firms but less than a third of the movers and eliminates roughly 30% of all workers regardless of their mobility status. These additional restrictions raise mean wages by roughly 5% and lower the variance of wages by 5–10% depending on the sample.

9.2. *Variance Decompositions*

We fit AKM models of the form given in (5) with $x_{gt} = 0$ to the leave-one-out samples after having pre-adjusted log wages for year effects in a first step.[10] The bottom of Table I reports for each sample the maximum leverage ($\max_i P_{ii}$) of any person-year observation.

[10]This adjustment is obtained by estimating an AKM model that includes a dummy control for the year 2001. Hence, y_{gt} gives the log wage in year t minus a year 2001 dummy times its estimated coefficient. This two-step

TABLE II

VARIANCE DECOMPOSITION[a]

	Pooled	Younger Workers	Older Workers
Variance of Firm Effects			
Plug in (PI)	0.0358	0.0368	0.0415
Homoscedasticity Only (HO)	0.0295	0.0270	0.0350
Leave Out (KSS)	0.0240	0.0218	0.0204
Variance of Person Effects			
Plug in (PI)	0.1321	0.0843	0.2180
Homoscedasticity Only (HO)	0.1173	0.0647	0.2046
Leave Out (KSS)	0.1119	0.0596	0.1910
Covariance of Firm, Person Effects			
Plug in (PI)	0.0039	−0.0058	−0.0032
Homoscedasticity Only (HO)	0.0097	0.0030	0.0040
Leave Out (KSS)	0.0147	0.0075	0.0171
Correlation of Firm, Person Effects			
Plug in (PI)	0.0565	−0.1040	−0.0334
Homoscedasticity Only (HO)	0.1649	0.0726	0.0475
Leave Out (KSS)	0.2830	0.2092	0.2744
Coefficient of Determination (R^2)			
Plug in (PI)	0.9546	0.9183	0.9774
Homoscedasticity Only (HO)	0.9029	0.8184	0.9524
Leave Out (KSS)	0.8976	0.8091	0.9489

[a] Decompositions conducted in the leave-one-out samples described in Table I. All variance components are person-year weighted. Wages have been pre-adjusted for a year fixed effect.

While our pruning procedure ensures $\max_i P_{ii} < 1$, it is noteworthy that $\max_i P_{ii}$ is still quite close to 1, indicating that certain person-year observations remain influential on the parameter estimates. This finding highlights the inadequacy of asymptotic approximations that require the dimensionality of regressors to grow slower than the sample size.

Table II reports the results of applying to our leave-one-out samples three estimators of the AKM variance decomposition: the traditional plug-in (PI) estimator $\hat{\theta}_{\text{PI}}$, the homoscedasticity-only (HO) estimator $\hat{\theta}_{\text{HO}}$ of Andrews et al. (2008), and the leave-out (KSS) estimator $\hat{\theta}$. The PI estimator finds that the variance of firm effects in the pooled sample accounts for roughly 20% of the total variance of wages, while among younger workers, firm effect variability is found to account for 31% of overall wage variance. Among older workers, variability in firm effects is estimated to account for only 16% of the variance of wages.

Applying the HO estimator of Andrews et al. (2008) reduces the estimated variances of firm effects by roughly 18% in the age-pooled sample, 27% in the sample of younger workers, and 16% in the sample of older workers. However, the KSS estimator yields further, comparably sized, reductions in the estimated firm effect variance relative to the HO estimator, indicating the presence of substantial heteroscedasticity in these samples. For instance, in the pooled leave-one-out sample, the KSS estimator finds a variance of firm effects that accounts for only 13% of the overall variance of wages, while the HO estimator finds that firm effects account for 16% of wage variance. Moreover, while the

approach simplifies computation without compromising consistency because the year effect is estimated at a $\sqrt{N}$ rate.

plug-in estimates suggested that the firm effect variance was greater among older than younger workers, the KSS estimator finds the opposite pattern.

PI estimates of person effect variances account for 66%–88% of the total variance of wages depending on the sample. Moreover, the estimated ratio of older to younger person effect variances in the leave-one-out sample is roughly 2.6. Applying the HO estimator reduces the magnitude of the person effect variance among all age groups, but boosts the ratio of older to younger person effect variances to 3.2. The KSS estimator yields further downward corrections to estimated person effect variances, leading the contribution of person effect variability to range from only 50% to 80% of total wage variance. Proportionally, however, the variability of older workers remains stable at 3.2 times that of younger workers.

PI estimates of the covariance between worker and firm effects are negative in both age-restricted samples, though not in the pooled sample. When converted to correlations, these figures suggest there is mild negative assortative matching of workers to firms. Applying the HO estimator leads the covariances to change sign in both age-specific samples, while generating a mild increase in the estimated covariance of the pooled sample. In all three samples, however, the HO estimates indicate very small correlations between worker and firm effects. By contrast, the KSS estimator finds a rather strong positive correlation of 0.21 among younger workers, 0.27 among older workers, and 0.28 in the pooled sample, indicating the presence of non-trivial positive assortative matching between workers and firms.

The usual PI estimator of R^2 suggests the two-way fixed effects model explains more than 95% of wage variation in the pooled sample, 91% in the sample of younger workers, and 97% in the sample of older workers. The HO estimator of R^2, which is equivalent to the adjusted R^2 measure of Theil (1961), indicates that the two-way fixed effects model explains roughly 90% of the variance of wages in the pooled sample, a finding quite close to the estimates reported in Card, Heining, and Kline (2013). Applying the heteroscedasticity robust KSS estimator yields negligible changes in estimated explanatory power relative to the HO estimates.[11] Interestingly, a sample size weighted average of the age group-specific KSS R^2 estimates lies slightly below the pooled KSS estimate of R^2, which suggests that allowing firm effects to differ by age group may fail to improve the model's fit. We examine this hypothesis more carefully in Section 9.3.

In Appendix A, we conduct variance decompositions in an unbalanced six year panel of VHW data and account for serial correlation in worker wages by leaving out "clusters" of wage observations as in Remark 3.[12] Remarkably, we find that leaving out either a worker-firm match or an entire worker history yields an estimated variance of firm effects very close to the two-period KSS estimates reported in Table II, suggesting that little serial correlation is present across worker-firm matches. Hence, a computationally convenient approach to avoiding biases in longer panels may be to simply collapse the data to match means in a first step and then analyze these means using the standard leave-one-observation-out estimator.

[11]A closed form expression for the KSS estimator of R^2, and its connection to the PI and HO estimators, was provided in an earlier version of this paper (Kline, Saggio, and Sølvsten (2019a)).

[12]Kline, Saggio, and Sølvsten (2019a) provided a more extensive analysis of this longer panel. Further analysis of KSS corrections in short and long panels can be found in Lachowska, Mas, Saggio, and Woodbury (2020).

TABLE III

PROJECTING FIRM EFFECTS ONTO COVARIATES[a]

	(1)	(2)
Older Worker	0.0272	−0.0016
	(0.0009)	(0.0024)
	[0.0003]	[0.0001]
Log Firm Size		0.0276
		(0.0007)
		[0.0001]
Older Worker × Log Firm Size		0.0028
		(0.0005)
		[0.0002]
Predicted Gap in Firm Effects (Older vs. Younger Workers)	0.0272	0.0054
	(0.0009)	(0.0019)
	[0.0003]	[0.0008]
Number of Observations	1,319,972	1,319,972

[a] This table reports the coefficients from projections of firm effects onto worker and firm characteristics in the pooled leave-one-out sample. A constant is included in each model. Standard errors based on equation (7) reported in parentheses. Naive Eicker–White (HC1) standard errors shown in square brackets. "Predicted Gap in Firm Effects" reports the predicted difference in firm effects between older and younger workers according to either Column (1) or Column (2) evaluated at the median firm size of 12 workers.

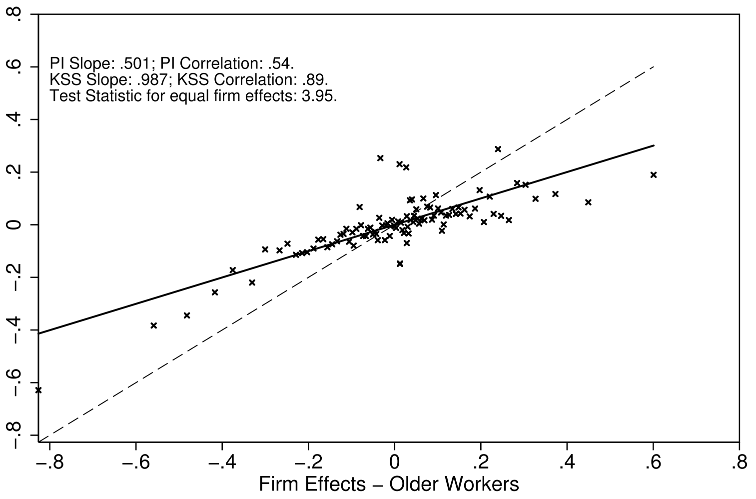
9.3. *Sorting and Wage Structure*

The KSS estimates reported in Table II suggest that older workers exhibit slightly less variable firm effects and a stronger correlation between person and firm effects than younger workers. These findings might reflect life cycle differences in the sorting of workers to firms or differences in the structure of firm wage effects across the two age groups.

Table III explores the sorting channel by projecting the pooled firm effects from the leave-one-out sample onto an indicator for being an older worker, the log of firm size, and their interaction. Because these projection coefficients are linear combinations of the estimated firm effects, we employ the KSS standard errors proposed in equation (7). For comparison, we also report naive Eicker–White standard errors that treat the firm effect estimates as mutually independent. In all cases, the KSS standard error is at least twice the corresponding naive standard error. Evidently, the standard practice of regressing fixed effect estimates on observables in a second step without accounting for correlation across these estimates can yield highly misleading inferences.

The first column of Table III shows that older workers tend to be employed at firms with firm wage effects roughly 2.7% higher than younger workers. The second column reveals that this sorting relationship is largely mediated by firm size. An older worker at a firm with a single employee is estimated to have a mean firm wage effect only 0.16% lower than a younger worker at a firm of the same size, an economically insignificant difference that is also deemed statistically insignificant when using the KSS standard error. As firm size grows, older workers begin to enjoy somewhat larger firm wage premia. At the median firm size of 12 workers, the predicted gap between older and younger workers rises to 0.54%, a modest gap that we can nonetheless distinguish from zero at the 5% level using the KSS standard error. We conclude that the tendency of older workers to be employed at larger firms is a quantitatively important driver of the firm wage premia they enjoy.

To assess whether firm wage effects differ between age groups, we study the set $\mathcal{J}$ of 8578 firms at which firm effects for younger (ψ_j^{Y}) and older (ψ_j^{O}) workers are both leave-one-out estimable. Figure 1 plots the person-year weighted averages of $\hat{\psi}_j^{\mathrm{Y}}$ and $\hat{\psi}_j^{\mathrm{O}}$ within



Notes: This figure plots means of the estimated firm effects for young workers ($\hat{\psi}_j^{\mathrm{Y}}$) against means of the estimated firm effects for older workers ($\hat{\psi}_j^{\mathrm{O}}$). Both sets of firm effect estimates have been demeaned. "PI slope" gives the coefficient from a person-year weighted projection of $\hat{\psi}_j^{\mathrm{Y}}$ onto $\hat{\psi}_j^{\mathrm{O}}$. "KSS slope" adjusts for attenuation bias by multiplying the PI slope by the ratio of the plug-in estimate of the person-year weighted variance of ψ_j^{O} to the KSS estimate of the same quantity. "PI correlation" gives the person-year weighted sample correlation between $\hat{\psi}_j^{\mathrm{O}}$ and $\hat{\psi}_j^{\mathrm{Y}}$, while "KSS correlation" adjusts this correlation for sampling error in both $\hat{\psi}_j^{\mathrm{O}}$ and $\hat{\psi}_j^{\mathrm{Y}}$ using KSS estimates of the relevant variances. "Test statistic" refers to the realization of $\hat{\theta}_{H_0}/\hat{\mathbb{V}}[\hat{\theta}_{H_0}]^{1/2}$ where $\hat{\theta}_{H_0}$ is the quadratic form associated with the null hypothesis that $\psi_j^{\mathrm{O}} = \psi_j^{\mathrm{Y}}$ for all 8,578 firms.

FIGURE 1.—Do firm effects differ across age groups?

each centile bin of $\hat{\psi}_j^{\mathrm{O}}$. A person-year weighted projection of $\hat{\psi}_j^{\mathrm{Y}}$ onto $\hat{\psi}_j^{\mathrm{O}}$ yields a slope of only 0.501. To correct for attenuation bias, we multiply this plug-in slope by the ratio of the PI estimate of the person-year weighted variance of ψ_j^{O} to the corresponding KSS estimate of this quantity, which yields an adjusted projection slope of 0.987. Converting this slope into a correlation using the KSS estimate of the person-year weighted variance of ψ_j^{Y} yields a person-year weighted correlation between the two sets of firm effects of 0.89. This high correlation suggests that the underlying $(\psi_j^{\mathrm{Y}}, \psi_j^{\mathrm{O}})$ pairs are in fact tightly clustered around the 45 degree line depicted in Figure 1.

Theorem 2 allows us to formally test the joint null hypothesis that the two sets of firm effects are actually identical; that is, that both the slope and R^2 from a projection of ψ_j^{Y} onto ψ_j^{O} equal 1. We can state this hypothesis as $H_0 : \psi_j^{\mathrm{O}} = \psi_j^{\mathrm{Y}}$ for all $j \in \mathcal{J}$. Using the test suggested in Remark 6, we obtain a realized test statistic of 3.95, which yields a p-value on H_0 of less than 0.1%. Hence, we can decisively reject the null hypothesis that older and younger workers face exactly the same vectors of firm effects, despite their high correlation.

9.4. *Inference*

We now study more carefully the problem of inference on the variance of firm effects. For convenience, the top row of Table IV reprints our earlier KSS estimates of the variance of firm effects in each sample. Below each estimate of firm effect variance is a standard error, computed according to the approach described in Lemma 5. As noted in Remark 8, these standard errors will be somewhat conservative when there is a large share of observations for which no split sample predictions can be created. In the leave-one-out samples, this share varies between 15% and 22%. For comparison, we also report results for leave-two-out samples, which turn out to exhibit very similar point estimates, and for which the split sample predictions always exist.[13] The standard errors will also tend to be conservative when there is a large share of observation pairs in the set $\mathcal{B}$, for which there is upward bias in the estimator of the error variance product. However, for both the leave-one-out and leave-two-out samples, this share varies between only 0.03% and 0.46%, suggesting that little bias stems from this source.

The next panel of Table IV reports the 95% confidence intervals that arise from setting $q = 0$, $q = 1$, or $q = 2$. While the first interval employs a normal approximation, the latter two allow for weak identification by employing non-standard limiting distributions involving linear combinations of normal and χ^2 random variables. We also report estimates of the curvature parameters (κ_1, κ_2) used to construct the weak identification robust intervals. In the pooled samples, both curvature parameters are estimated to be quite small, indicating that a normal approximation is accurate. Accordingly, setting $q > 0$ has little discernible effect on the resulting confidence intervals in these samples. However, among older workers, particularly in the leave-two-out sample, we find greater curvature, suggesting weak identification may be empirically relevant. Setting $q > 0$ in this sample widens the confidence interval somewhat and also changes its shape: mildly shortening the lower tail of the interval but lengthening the upper tail.

Theorem 3 suggests that two important diagnostics for the asymptotic behavior of our estimator are the top eigenvalue shares $\{\lambda_s^2 / \sum_{\ell=1}^{r} \lambda_\ell^2\}_{s=1,2,3}$ and the Lindeberg statistics $\{\max_i \mathrm{w}_{is}^2\}_{s=1,2}$. The bottom panel of Table IV reports these statistics for each sample. The top eigenvalue shares are fairly small in the pooled sample and among younger workers. A small top eigenvalue share indicates that the estimator does not depend strongly on any particular linear combination of firm effects and hence that a normal distribution should provide a suitable approximation to the estimator's asymptotic behavior (i.e., that $q = 0$). Accordingly, we find that the confidence intervals are virtually identical for all values of q in both the pooled samples and the two samples of younger workers.

Among older workers, the top eigenvalue share is 31% in the leave-one-out sample and 58% in the leave-two-out sample. The next largest eigenvalue share is, in both cases, less than 5%, which suggests this is a setting where $q = 1$. In line with the theory, confidence intervals based upon the $q = 1$ and $q = 2$ approximations are nearly identical in both samples of older workers. The accuracy of these weak-identification robust confidence intervals hinges on the Lindeberg condition of Theorem 3 being satisfied. One can think of the Lindeberg statistic $\max_i \mathrm{w}_{is}^2$ as giving an inverse measure of effective sample size available for estimating the linear combination of firm effects associated with the sth largest eigenvalue. The fact that these statistics are all less than or equal to 0.05 implies an effective sample size of at least 20. Finally, the sum of squared eigenvalues is quite small in all six samples considered, indicating that the leave-out estimator is consistent also in our weakly identified samples.

9.5. *Monte Carlo Experiments*

We turn now to studying the finite sample behavior of the leave-out estimator of firm effect variance and its associated confidence intervals under a particular data generating

[13]See Kline, Saggio, and Sølvsten (2019a) for summary statistics on this sample and a more detailed comparison of leave-one-out and leave-two-out estimates.

TABLE IV

INFERENCE ON THE VARIANCE OF FIRM EFFECTS[a]

	Pooled		Younger Workers		Older Workers	
	Leave-One-Out Sample	Leave-Two-Out Sample	Leave-One-Out Sample	Leave-Two-Out Sample	Leave-One-Out Sample	Leave-Two-Out Sample
Variance of Firm Effects						
Leave-out estimate	0.0240	0.0238	0.0218	0.0221	0.0204	0.0180
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0025)	(0.0013)
Sum of Squared Eigenvalues	2.11×10^{-6}	1.62×10^{-6}	3.97×10^{-6}	2.99×10^{-6}	0.0001	0.0001
Percentage of movers with no split sample estimator	14.92%	0.00%	21.00%	0.00%	22.00%	0.00%
Percentage of mover pairs in $\mathcal{B}$	0.04%	0.05%	0.03%	0.05%	0.21%	0.46%
95% *Confidence Intervals*						
Strong id ($q=0$)	[0.0228; 0.0251]	[0.0227; 0.0249]	[0.0207; 0.0230]	[0.0210; 0.0232]	[0.0155; 0.0254]	[0.0154; 0.0205]
Weak id ($q=1$)	[0.0228; 0.0251]	[0.0227; 0.0249]	[0.0207; 0.0230]	[0.0209; 0.0234]	[0.0157; 0.0267]	[0.0158; 0.0221]
Weak id ($q=2$)	[0.0228; 0.0251]	[0.0227; 0.0251]	[0.0207; 0.0231]	[0.0209; 0.0234]	[0.0158; 0.0272]	[0.0161; 0.0231]
Curvature ($\kappa_1, q=1$)	0.0182	0.0416	0.0197	0.0689	0.3333	0.7845
Curvature ($\kappa_2, q=2$)	0.0271	0.1458	0.0444	0.0678	0.2779	1.1088
Diagnostics						
Lindeberg Condition ($q=1$)	0.1878	0.0865	0.2061	0.0371	0.0359	0.0503
Lindeberg Condition ($q=2$)	0.0866	0.1604	0.0237	0.1639	0.0491	0.0479
Eigenvalue Ratio—1	0.0135	0.0233	0.0189	0.0211	0.3132	0.5883
Eigenvalue Ratio—2	0.0131	0.0202	0.0058	0.0135	0.0387	0.0499
Eigenvalue Ratio—3	0.0112	0.0131	0.0050	0.0087	0.0314	0.0187

[a]This table conducts inference on the variance of firm effects using the samples described in Table I. The parentheses report standard errors as described in Section 5.2. Confidence intervals are computed under different assumptions on q (see Section 7.1). "Curvature" reports the maximal curvature; see the Supplemental Material for further details. "Eigenvalue ratio—s" gives the ratio of the square of the sth largest eigenvalue of the matrix $\tilde{A}$ to the sum of all its squared eigenvalues. "Percentage of movers with no split sample estimator" reports the percentage of movers for which it is impossible to find two independent unbiased estimators of their conditional mean. "Percentage of mover pairs in $\mathcal{B}$" gives the fraction of mover pairs where we use an unconditional variance estimate (see Section 5.2).

TABLE V

MONTE CARLO RESULTS FOR THE VARIANCE OF FIRM EFFECTS[a]

	Pooled		Younger Workers		Older Workers	
	Leave-One-Out Sample	Leave-Two-Out Sample	Leave-One-Out Sample	Leave-Two-Out Sample	Leave-One-Out Sample	Leave-Two-Out Sample
Relative Bias in Point Estimators						
Leave Out (KSS)	0.03%	0.02%	−0.04%	−0.02%	0.41%	−0.44%
	(1.15%)	(1.35%)	(1.37%)	(1.57%)	(7.23%)	(10.61%)
Homoscedasticity Only (HO)	24.12%	15.02%	29.93%	18.06%	33.63%	19.54%
	(1.13%)	(1.35%)	(1.36%)	(1.59%)	(7.25%)	(10.60%)
Plug in (PI)	28.19%	17.94%	36.67%	22.58%	37.55%	22.01%
	(1.13%)	(1.35%)	(1.36%)	(1.59%)	(7.25%)	(10.60%)
Relative Bias in KSS Std Error						
	41.04%	3.42%	47.97%	3.99%	27.92%	−0.33%
Coverage rate						
Strong id ($q = 0$)	99.6%	96.4%	99.5%	95.4%	97.9%	89.9%
Weak id ($q = 1$)	99.6%	96.3%	99.3%	96.5%	98.5%	95.5%

[a]Section 9.5 describes the DGP. "Relative Bias in Point Estimators" gives the average across simulations of the difference between the estimated and true values of the firm effect variance scaled by the true variance of firm effects. In parentheses, the table reports the simulated standard deviation of the same quantity. "Relative Bias in KSS Std Error" reports the average across simulations of the difference between the KSS standard error and the Monte Carlo standard deviation of the KSS estimator scaled by the Monte Carlo standard deviation of the KSS estimator. "Strong id" gives the coverage rate of a confidence interval for the variance of firms effects based upon KSS standard errors and a normal approximation. "Weak id" reports the coverage rate of the test-inversion based confidence interval described in Section 7 under $q = 1$. All results rely upon 1000 Monte Carlo draws.

process (DGP). Data were generated from the following first-differenced model based upon equation (6):

$$\Delta y_g = \Delta f_g' \hat{\psi}^{\text{scale}} + \Delta \varepsilon_g \quad (g = 1, \dots, N).$$

Here, $\hat{\psi}^{\text{scale}}$ gives the vector of OLS firm effect estimates found in the pooled leave-one-out sample, rescaled to match the KSS estimate of firm effect variance for that sample of 0.024. The errors $\Delta\varepsilon_g$ were drawn independently from a Student's t-distribution with 5 degrees of freedom and variances given by the following model:

$$\mathbb{V}[\Delta\varepsilon_g] = \exp(a_0 + a_1 B_{gg} + a_2 P_{gg} + a_3 \ln L_{g2} + a_4 \ln L_{g1}),$$

where L_{gt} gives the size of the firm employing worker g in period t. The coefficients of this model were estimated via a nonlinear least squares fit to the $\hat{\sigma}_g^2$ in the pooled leave-one-out sample.[14] For each sample, we drew from the above DGP 1000 times while holding firm assignments fixed at their sample values.

Table V reports the results of this Monte Carlo experiment. In accord with theory, the KSS estimator of firm effect variances is unbiased while the PI and HO estimators are biased upwards. As expected, the KSS standard error estimator exhibits a modest upward bias in the leave-one-out samples ranging from 28% in the sample of older workers to 48% among younger workers. In the leave-two-out sample, however, the standard error estimator exhibits biases of only 4% or less. Unsurprisingly then, the $q = 0$ confidence interval over-covers in both the pooled leave-one-out sample and the leave-one-out sample

[14]The parameter estimates were: $\hat{a}_0 = -3.3441$, $\hat{a}_1 = 1.3951$, $\hat{a}_2 = -0.0037$, $\hat{a}_3 = -0.0012$, $\hat{a}_4 = -0.0086$.

of younger workers. In the corresponding leave-two-out samples, however, coverage is very near its nominal level, both for the normal based ($q = 0$) and the weak identification robust ($q = 1$) intervals.

In the samples of older workers, the normal distribution provides a poor approximation to the shape of the estimator's sampling distribution, which is to be expected given the large top eigenvalues found in these designs. This non-normality leads to under-coverage by the $q = 0$ confidence interval in the leave-two-out sample. By contrast, applying the weak identification robust interval yields coverage very close to nominal levels despite the fact that the effective sample size available for the top eigenvector is only about 20.

10. CONCLUSION

We propose a new estimator of quadratic forms with applications to several areas of economics. The estimator is finite sample unbiased in the presence of unrestricted heteroscedasticity and can be accurately approximated in very large data sets via random projection methods. Consistency is established under verifiable design requirements in an environment where the number of regressors may grow in proportion to the sample size. The estimator enables tests of linear restrictions of varying dimension under weaker conditions than have been explored in previous work. A new distributional theory highlights the potential for the proposed estimator to exhibit deviations from normality when some linear combinations of coefficients are imprecisely estimated relative to others. Monte Carlo experiments demonstrate that confidence intervals predicated on the assumption that $q = 1$ can provide accurate size control, even when the realized mobility network exhibits a severe bottleneck.

APPENDIX A: ANALYSIS OF LONGER PANELS

This appendix reports estimates of the variance of firm effects in an unbalanced panel spanning the years 1996–2001.[15] Because the equivalence discussed in Remark 4 no longer holds when $T > 2$, the leave-out estimator may exhibit a bias when the errors are serially correlated. Table A.I probes for the importance of serial correlation by leaving out "clusters" of observations—as described in Remark 3—defined successively as all observations within the same worker-firm "match" and all observations belonging to the same worker.[16]

Leaving out the match yields an important reduction in the variance of firm effects relative to leaving out a single person-year observation, indicating the presence of substantial serial correlation within match. By contrast, leaving out all observations associated with the worker turns out to have negligible effects on the estimated variance of firm effects, suggesting that serial correlation across matches is negligible. As expected, pooling several years of data reduces the bias of the plug-in estimator. Contrasting Table A.I with Table II, the leave-worker-out bias correction reduces the pooled plug-in estimate of firm effect variance in our unbalanced six-year panel by roughly 20%, while bias correcting the corresponding plug-in estimate in our baseline two-year sample yields a reduction of approximately 30%.

[15]To analyze this longer panel, we expand our set of time varying covariates to include unrestricted year effects and a third-order polynomial in age normalized to have slope zero at age 40 as discussed in Card, Cardoso, Heining, and Kline (2018). Pre-adjusting for age has negligible effects on the variance decompositions reported in Table II but is quantitatively more important in this longer panel.

[16]Because worker g's person effect is not estimable when leaving that worker's entire wage history out, we estimate within-transformed specifications that eliminate the person effects in a first step.

TABLE A.I

VARIANCE OF FIRM EFFECTS UNDER DIFFERENT LEAVE-OUT STRATEGIES[a]

	Pooled	Younger Workers	Older Workers
Variance of Firm Effects			
Plug-in	0.0304	0.0303	0.0376
Leave Person-Year Out	0.0296	0.0302	0.0314
Leave Match Out	0.0243	0.0221	0.0265
Leave Worker Out	0.0241	0.0227	0.0270

[a]Decompositions use all VHW wage records spanning the years 1996–2001 that are leave-worker-out connected in the relevant sample. The pooled sample contains 5,163,446 person-year observations, the sample of younger workers contains 2,632,596 observations, and the sample of older workers contains 2,016,202 observations; see Table 3 of Kline, Saggio, and Sølvsten (2019a) for additional sample dimensions and summary statistics. "Leave Person-Year Out" computes a bias correction by leaving a single person-year observation out. "Leave Match Out" computes a bias correction by leaving entire worker-firm matches out. "Leave Worker Out" computes a bias correction by leaving out each worker's entire wage history. The Computational Appendix provides implementation details.

APPENDIX B: PROOFS

PROOF OF LEMMA 1: For the first claim, it suffices to show that $\mathbb{E}[\hat{\sigma}_i^2] = \sigma_i^2$ when $P_{ii} < 1$, since $\hat{\theta} = \hat{\beta}' A \hat{\beta} - \sum_{i=1}^n B_{ii}\hat{\sigma}_i^2$ and $\mathbb{E}[\hat{\beta}' A \hat{\beta}] - \theta = \text{trace}(A\mathbb{V}[\hat{\beta}]) = \sum_{i=1}^n B_{ii}\sigma_i^2$. When S_{xx} has full rank and $P_{ii} < 1$, it follows from the Sherman–Morrison–Woodbury formula that $S_{xx} - x_i x_i'$ is invertible so that the leave-one-out estimator $\hat{\beta}_{-i} = (S_{xx} - x_i x_i')^{-1} \sum_{\ell \neq i} x_\ell y_\ell$ exists. As $\hat{\beta}_{-i}$ is independent of ε_i and unbiased for β with fixed regressors, we have

$$\begin{aligned}
\mathbb{E}\big[\hat{\sigma}_i^2\big] &= \mathbb{E}\big[y_i\big(y_i - x_i'\hat{\beta}_{-i}\big)\big] = \mathbb{E}\big[\big(\varepsilon_i + x_i'\beta\big)(\varepsilon_i + x_i'(\beta - \hat{\beta}_{-i})\big] \\
&= \mathbb{E}\big[\varepsilon_i^2\big] + \mathbb{E}[\varepsilon_i]x_i'\mathbb{E}[\beta - \hat{\beta}_{-i}] + x_i'\beta\mathbb{E}[\varepsilon_i] + x_i'\beta x_i'\mathbb{E}[\beta - \hat{\beta}_{-i}] = \sigma_i^2.
\end{aligned}$$

For the second claim, it suffices to show that no unbiased estimator of $\beta' S_{xx} \beta$ exists when $\max_i P_{ii} = 1$. As the model only places restrictions on the first two moments of y_i, any unbiased estimator must have the form $y'\mathcal{C}y + U$, where $y = (y_1, \dots, y_n)'$, $\mathbb{E}[U] = 0$, and $\mathcal{C} = (\mathcal{C}_{i\ell})_{i,\ell}$ satisfies (i) $\mathcal{C}_{ii} = 0$ for all i and (ii) $X'\mathcal{C}X = S_{xx}$ for $X = (x_1, \dots, x_n)'$. Item (ii) implies that $\mathcal{C}$ must satisfy $\mathcal{C} = I + P\tilde{\mathcal{C}}M + M\tilde{\mathcal{C}}P + M\tilde{\mathcal{C}}M$ for some $\tilde{\mathcal{C}}$ where $M = (M_{i\ell})_{i,\ell}$ and $P = I_n - M$. If there exists an i with $P_{ii} = 1$, then $\sum_{\ell=1}^n P_{i\ell}^2 = P_{ii}$ yields $M_{i\ell} = 0$ for all ℓ which implies that $\mathcal{C}_{ii}$ must equal 1 to satisfy (ii). However, this makes it impossible to satisfy (i), so no unbiased estimator can exist. *Q.E.D.*

PROOF OF LEMMA 2: Recall the spectral decomposition $\tilde{A} = QDQ'$ and definition of $\hat{b} = Q' S_{xx}^{1/2} \hat{\beta}$ which satisfies that $\hat{b} \sim \mathcal{N}(b, \mathbb{V}[\hat{b}])$ when $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. We have that $\theta^* = \sum_{\ell=1}^r \lambda_\ell(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell])$ since $\hat{\beta}' A \hat{\beta} = \hat{\beta}' S_{xx}^{1/2} \tilde{A} S_{xx}^{1/2} \hat{\beta} = \hat{b}' D \hat{b} = \sum_{\ell=1}^r \lambda_\ell \hat{b}_\ell^2$ and $\sum_{i=1}^n B_{ii}\sigma_i^2 = \text{trace}(A\mathbb{V}[\hat{\beta}]) = \text{trace}(D\mathbb{V}[\hat{b}]) = \sum_{\ell=1}^r \lambda_\ell \mathbb{V}[\hat{b}_\ell]$. *Q.E.D.*

PROOF OF LEMMA 3: The difference between $\hat{\theta}$ and θ is

$$\hat{\theta} - \theta = 2\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x_\ell' \beta \varepsilon_i + \sum_{i=1}^n \sum_{\ell \neq i} B_{i\ell}\varepsilon_i\varepsilon_\ell + \sum_{i=1}^n B_{ii}\big(\varepsilon_i^2 - \hat{\sigma}_i^2\big),$$

and each term has mean zero so we show that their variances are small in large samples. Suppose that A is positive semi-definite. The variance of the first term is

$$4\sum_{i=1}^{n}\left(\sum_{\ell=1}^{n}B_{i\ell}x_\ell'\beta\right)^2\sigma_i^2 \le 4\max_i\sigma_i^2\beta'X'B^2X\beta = 4\max_i\sigma_i^2\beta'AS_{xx}^{-1}A\beta$$
$$\le 4\max_i\sigma_i^2\theta\lambda_1 = o(1),$$

where $B=(B_{i\ell})_{i,\ell}$, the last inequality follows from positive semi-definiteness of A, and the last equality follows from $\theta = O(1)$ and $\lambda_1 \le \text{trace}(\tilde{A}^2)^{1/2} = o(1)$. The variance of the second term is

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

Finally, the variance of the third term is

$$\sum_{i=1}^{n}\left(\sum_{\ell=1}^{n}M_{\ell\ell}^{-1}B_{\ell\ell}M_{i\ell}x_\ell'\beta\right)^2\sigma_i^2 + 2\sum_{i=1}^{n}\sum_{\ell\neq i}M_{ii}^{-2}B_{ii}^2M_{i\ell}^2\sigma_i^2\sigma_\ell^2$$
$$\le \frac{1}{c^2}\max_i\sigma_i^2\max_i(x_i'\beta)^2\sum_{i=1}^{n}B_{ii}^2 + \frac{2}{c}\max_i\sigma_i^4\sum_{i=1}^{n}B_{ii}^2 = o(1),$$

where $\min_i M_{ii} \ge c > 0$ and $\sum_{i=1}^{n}B_{ii}^2 \le \text{trace}(\tilde{A}^2) = o(1)$. This shows the first claim of the lemma.

When A is non-definite, we write $A = \frac{1}{2}(A_1'A_2 + A_2'A_1)$ and note that

$$\beta'AS_{xx}^{-1}A\beta \le \frac{1}{2}\big(\theta_1\lambda_{\max}(\tilde{A}_2) + \theta_2\lambda_{\max}(\tilde{A}_1)\big) \quad \text{and}$$
$$\text{trace}(\tilde{A}^2) \le \text{trace}(\tilde{A}_1^2)^{1/2}\,\text{trace}(\tilde{A}_2^2)^{1/2},$$

where $\tilde{A}_\ell = S_{xx}^{-1/2}A_\ell'A_\ell S_{xx}^{-1/2}$ for $\ell = 1, 2$ and $\lambda_{\max}(\tilde{A}_2)$ is the largest eigenvalue of $\tilde{A}_2$. Thus, consistency of $\hat{\theta}$ follows from $\theta_\ell = O(1)$ and $\text{trace}(\tilde{A}_\ell^2) = o(1)$ for $\ell = 1, 2$. *Q.E.D.*

PROOF OF LEMMA 4: See Supplemental Material. *Q.E.D.*

PROOF OF THEOREM 1: The proof has two steps. First, we write $\hat{\theta}$ as $\sum_{\ell=1}^{r}\lambda_\ell(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell])$ plus an approximation error of smaller order than $\mathbb{V}[\hat{\theta}]$. This argument establishes the last two claims of the lemma. Second, we use Lyapunov's CLT to show that $\hat{b} \in \mathbb{R}^r$ is jointly asymptotically normal.

Decomposition and Approximation. From the proof of Lemma 2, it follows that

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

and we show next that the mean-zero variable $\sum_{i=1}^n B_{ii}(\sigma_i^2 - \hat{\sigma}_i^2)$ is $o_p(\mathbb{V}[\hat{\theta}]^{1/2})$. We have

$$\sum_{i=1}^n B_{ii}\big(\hat{\sigma}_i^2 - \sigma_i^2\big) = \sum_{i=1}^n B_{ii} \sum_{\ell=1}^n M_{ii}^{-1} x_i'\beta M_{i\ell}\varepsilon_\ell + \sum_{i=1}^n B_{ii}\big(\varepsilon_i^2 - \sigma_i^2\big) + \sum_{i=1}^n B_{ii} \sum_{\ell\neq i} M_{ii}^{-1} M_{i\ell}\varepsilon_i\varepsilon_\ell.$$

The variances of these three terms are

$$\begin{aligned}
&\sum_{\ell=1}^n \sigma_\ell^2 \left(\sum_{i=1}^n M_{i\ell} B_{ii} M_{ii}^{-1} x_i'\beta\right)^2 \\
&\qquad \le \max_i \sigma_i^2 \sum_{i=1}^n B_{ii}^2 M_{ii}^{-2}\big(x_i'\beta\big)^2 \le \max_i \sigma_i^2 \max_i \big(x_i'\beta\big)^2 M_{ii}^{-2} \times \sum_{i=1}^n B_{ii}^2, \\
&\sum_{i=1}^n B_{ii}^2 \mathbb{V}\big[\varepsilon_i^2\big] \le \max_i \mathbb{E}\big[\varepsilon_i^4\big] \times \sum_{i=1}^n B_{ii}^2, \\
&\sum_{i=1}^n \sum_{\ell\neq i} \big(B_{ii}^2 M_{ii}^{-2} + B_{ii} M_{ii}^{-1} B_{\ell\ell} M_{\ell\ell}^{-1}\big) M_{i\ell}^2 \sigma_i^2 \sigma_\ell^2 \le 2 \max_i \sigma_i^4 M_{ii}^{-2} \times \sum_{i=1}^n B_{ii}^2.
\end{aligned}$$

Furthermore, we have that

$$\mathbb{V}[\hat{\theta}]^{-1} \sum_{i=1}^n B_{ii}^2 \le \max_i w_i'w_i \mathbb{V}[\hat{\theta}]^{-1} \sum_{l=1}^r \lambda_l^2(\tilde{A}) \le \max_i w_i'w_i \max_i \sigma_i^{-4} = o(1),$$

so each of the three variances is of smaller order than $\mathbb{V}[\hat{\theta}]$.

For the second claim, it suffices to show that $\delta(v) := \mathbb{V}[v'\hat{b}]^{-1}(\hat{\mathbb{V}}[v'\hat{b}] - \mathbb{V}[v'\hat{b}]) = o_p(1)$ for all non-random $v \in \mathbb{R}^r$ with $v'v = 1$. Let $v \in \mathbb{R}^r$ be such a vector. As above, we have that $\delta(v) = \sum_{i=1}^n w_i(v)(\hat{\sigma}_i^2 - \sigma_i^2)$ is a mean-zero variable which is $o_p(1)$ if $\sum_{i=1}^n w_i(v)^4 = o(1)$ where $w_i(v) = (v'w_i)^2/\sum_{i=1}^n \sigma_i^2(v'w_i)^2$. But this follows from $\sum_{i=1}^n w_i(v)^4 \le \max_i \sigma_i^{-4} \max_i w_i'w_i = o(1)$ where the inequality is implied by $\max_i w_i'w_i = o(1)$, $v'v = 1$, and $\sum_{i=1}^n w_i w_i' = I_r$.

Asymptotic Normality. Next, we show that all linear combinations of $\hat{b}$ are asymptotically normal. Let $v \in \mathbb{R}^r$ be a non-random vector with $v'v = 1$. Lyapunov's CLT implies that $\mathbb{V}[v'\hat{b}]^{-1/2} v'(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, 1)$ if

$$\mathbb{V}\big[v'\hat{b}\big]^{-2} \sum_{i=1}^n \mathbb{E}\big[\varepsilon_i^4\big]\big(v'Q'S_{xx}^{-1/2} x_i\big)^4 = \mathbb{V}\big[v'\hat{b}\big]^{-2} \sum_{i=1}^n \mathbb{E}\big[\varepsilon_i^4\big]\big(v'w_i\big)^4 = o(1). \tag{9}$$

We have that $\max_i w_i'w_i = o(1)$ implies (9) since $\max_i(v'w_i)^2 \le \max_i w_i'w_i$, $\sum_{i=1}^n (v'w_i)^2 = 1$, $\mathbb{V}[v'\hat{b}]^{-1} \le \max_i \sigma_i^{-2} = O(1)$, and $\max_i \mathbb{E}[\varepsilon_i^4] = O(1)$ by the definition of w_i and Assumption 1. *Q.E.D.*

The proofs of Theorems 2 and 3 are based on the following lemma. Let $\{v_{n,i}\}_{i,n}$ be a triangular array of row-wise independent random variables with $\mathbb{E}[v_{n,i}] = 0$ and $\mathbb{V}[v_{n,i}] = \sigma_{n,i}^2$, let $\{\dot{w}_{n,i}\}_{i,n}$ be a triangular array of non-random weights that satisfy $\sum_{i=1}^n \dot{w}_{n,i}^2 \sigma_{n,i}^2 = 1$ for all n, and let $(W_n)_n$ be a sequence of symmetric non-random matrices in $\mathbb{R}^{n\times n}$ with zeroes on the diagonal that satisfy $2\sum_{i=1}^{n}\sum_{\ell\neq i}W_{n,i\ell}^2\sigma_{n,i}^2\sigma_{n,\ell}^2=1$. For simplicity, we drop the subscript n on $v_{n,i}$, $\sigma_{n,i}^2$, $\dot{w}_{n,i}$, $W_{n,i\ell}$, and W_n. Define

$$\mathcal{S}_n=\sum_{i=1}^{n}\dot{w}_iv_i \quad \text{and} \quad \mathcal{U}_n=\sum_{i=1}^{n}\sum_{\ell\neq i}W_{i\ell}v_iv_\ell.$$

LEMMA B.1: *If* $\max_i\mathbb{E}[v_i^4]+\sigma_i^{-2}=O(1)$, (i) $\max_i\dot{w}_i^2=o(1)$, *and* (ii) $\operatorname{trace}(W^4)=o(1)$, *then* $(\mathcal{S}_n,\mathcal{U}_n)'\xrightarrow{d}\mathcal{N}(0,I_2)$.

This lemma extends the main result of Appendix A2 in Sølvsten (2020) to allow for $\{v_i\}_i$ to be a triangular array of non-identically distributed variables. Furthermore, the conclusion is presented in a way that is tailored to the subsequent proofs in this paper. The proof of Lemma B.1 requires no substantially new ideas compared to Sølvsten (2020).

PROOF OF LEMMA B.1: See Supplemental Material. *Q.E.D.*

PROOF OF THEOREM 2: The proof involves two steps. First, we decompose $\hat{\theta}$ into a weighted sum of two terms of the type described in Lemma B.1. Second, we use Lemma B.1 to show joint asymptotic normality of the two terms. The conclusion that $\hat{\theta}$ is asymptotically normal is immediate from there.

Decomposition. The difference between $\hat{\theta}$ and θ is

$$\hat{\theta}-\theta=\sum_{i=1}^{n}\bigl(2\tilde{x}_i'\beta-\check{x}_i'\beta\bigr)\varepsilon_i+\sum_{i=1}^{n}\sum_{\ell\neq i}C_{i\ell}\varepsilon_i\varepsilon_\ell,$$

where these two terms are uncorrelated and have variances

$$V_{\mathcal{S}}=\sum_{i=1}^{n}\bigl(2\tilde{x}_i'\beta-\check{x}_i'\beta\bigr)^2\sigma_i^2 \quad \text{and} \quad V_{\mathcal{U}}=2\sum_{i=1}^{n}\sum_{\ell\neq i}C_{i\ell}^2\sigma_i^2\sigma_\ell^2.$$

Thus, we write $\mathbb{V}[\hat{\theta}]^{-1/2}(\hat{\theta}-\theta)=\omega_1\mathcal{S}_n+\omega_2\mathcal{U}_n$, where $\omega_1=V_{\mathcal{S}}^{1/2}/\mathbb{V}[\hat{\theta}]^{1/2}$, $\omega_2=V_{\mathcal{U}}^{1/2}/\mathbb{V}[\hat{\theta}]^{1/2}$,

$$\mathcal{S}_n=V_{\mathcal{S}}^{-1/2}\sum_{i=1}^{n}\bigl(2\tilde{x}_i'\beta-\check{x}_i'\beta\bigr)\varepsilon_i \quad \text{and} \quad \mathcal{U}_n=V_{\mathcal{U}}^{-1/2}\sum_{i=1}^{n}\sum_{\ell\neq i}C_{i\ell}\varepsilon_i\varepsilon_\ell.$$

Asymptotic Normality. We will argue along converging subsequences. Move to a subsequence where ω_1 converges. If the limit is zero, then $\mathbb{V}[\hat{\theta}]^{-1/2}(\hat{\theta}-\theta)=\omega_2\mathcal{U}_n+o_p(1)$ and so it follows from marginal normality of $\mathcal{U}_n$ established in the last paragraph of the proof that $\hat{\theta}$ is asymptotically normal. Thus, we consider the case where the limit of ω_1 is nonzero. In the notation of Lemma B.1, we then have $\dot{w}_i=V_{\mathcal{S}}^{-1/2}(2\tilde{x}_i'\beta-\check{x}_i'\beta)$ and $W_{i\ell}=V_{\mathcal{U}}^{-1/2}C_{i\ell}$.

For Lemma B.1(i), we have

$$\max_i\dot{w}_i^2\leq 4\omega_1^{-1}\max_i\frac{\bigl(\tilde{x}_i'\beta\bigr)^2+\bigl(\check{x}_i'\beta\bigr)^2}{\mathbb{V}[\hat{\theta}]}=o(1),$$

where the last equality follows from Theorem 2(i) and the nonzero limit of ω_1.

For Lemma B.1(ii), it can be shown that for all n, $\text{trace}(C^4) \le c_U \cdot \text{trace}(B^4) = c_U \cdot \text{trace}(\tilde{A}^4) \le c_U \lambda_1^2 \cdot \text{trace}(\tilde{A}^2)$ and $V_{\mathcal{U}} \ge c_L \min_i \sigma_i^4 \cdot \text{trace}(\tilde{A})$, where the finite and nonzero constants c_U and c_L do not depend on n (but depend on $\min_i M_{ii}$ which is bounded away from zero). Thus, Assumption 1 implies that

$$\text{trace}\big(W^4\big) \le \frac{c_U \lambda_1^2 \cdot \text{trace}\big(\tilde{A}^2\big)}{\big(c_L \min_i \sigma_i^4 \cdot \text{trace}\big(\tilde{A}^2\big)\big)^2} = O\left(\frac{\lambda_1^2}{\text{trace}\big(\tilde{A}^2\big)}\right) = o(1),$$

where the last equality follows from Theorem 2(ii). *Q.E.D.*

PROOF OF LEMMA 5: See Supplemental Material. *Q.E.D.*

Exact definitions of the variables involved in stating the regularity conditions of Theorem 3 were omitted from the main text and are provided here. Let $C_{i\ell q} = B_{i\ell q} - 2^{-1} M_{i\ell}(M_{ii}^{-1} B_{iiq} + M_{\ell\ell}^{-1} B_{\ell\ell q})$, $B_{i\ell q} = x_i' S_{xx}^{-1/2} \tilde{A}_q S_{xx}^{-1/2} x_\ell$, $\tilde{A}_q = \sum_{\ell=q+1}^{r} \lambda_\ell q_\ell q_\ell'$, $\tilde{x}_{iq} = \sum_{\ell=1}^{n} B_{i\ell q} x_\ell$, and $\check{x}_{iq} = \sum_{\ell=1}^{n} M_{i\ell} M_{\ell\ell}^{-1} B_{\ell\ell q} x_\ell$.

PROOF OF THEOREM 3: The proof involves two steps. First, we write $\hat{\theta}$ as the sum of (1a) a quadratic function applied to $\hat{\mathsf{b}}_q$, (1b) an approximation error which is of smaller order than $\mathbb{V}[\hat{\theta}]$, and (2) a weighted sum of two terms, $\mathcal{S}_n$ and $\mathcal{U}_n$, of the type described in Lemma B.1. Second, we use Lemma B.1 to show that $(\hat{\mathsf{b}}_q', \mathcal{S}_n, \mathcal{U}_n)' \in \mathbb{R}^{q+2}$ is jointly asymptotically normal.

Decomposition and Approximation. Noting that $\hat{\beta}' A \hat{\beta} = \sum_{\ell=1}^{q} \lambda_\ell \hat{b}_\ell^2 + \sum_{i=1}^{n} \sum_{\ell=1}^{n} B_{i\ell q} y_i y_\ell$ and

$$\begin{aligned}\sum_{i=1}^{n} B_{ii} \hat{\sigma}_i^2 &= \sum_{i=1}^{n} B_{ii,-q} \sigma_i^2 + \sum_{i=1}^{n} B_{iiq} \hat{\sigma}_i^2 + \sum_{i=1}^{n} B_{ii,-q}\big(\hat{\sigma}_i^2 - \sigma_i^2\big) \\ &= \sum_{\ell=1}^{q} \lambda_\ell \mathbb{V}[\hat{b}_\ell] + \sum_{i=1}^{n} B_{iiq} \hat{\sigma}_i^2 + o_p\big(\mathbb{V}[\hat{\theta}]^{1/2}\big) \quad \text{where } B_{ii,-q} = B_{ii} - B_{iiq},\end{aligned}$$

we have that

$$\hat{\theta} = \sum_{\ell=1}^{q} \lambda_\ell \big(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell]\big) + \hat{\theta}_q + o_p\big(\mathbb{V}[\hat{\theta}]^{1/2}\big) \quad \text{for } \hat{\theta}_q = \sum_{i=1}^{n} \sum_{\ell \ne i} C_{i\ell q} y_i y_\ell,$$

where it follows from $\max_i \mathsf{w}_{iq}' \mathsf{w}_{iq} = o(1)$ and the calculations in the proof of Theorem 1 that the mean-zero random variable $\sum_{i=1}^{n} B_{ii,-q}(\hat{\sigma}_i^2 - \sigma_i^2)$ is $o_p(\mathbb{V}[\hat{\theta}]^{1/2})$.

We will further center and rescale $\hat{\theta}_q$ by writing

$$\mathbb{V}[\hat{\theta}_q]^{-1/2}\big(\hat{\theta}_q - \mathbb{E}[\hat{\theta}_q]\big) = \omega_1 \mathcal{S}_n + \omega_2 \mathcal{U}_n,$$

where $\omega_1 = V_{\mathcal{S}}^{1/2} / \mathbb{V}[\hat{\theta}_q]^{1/2}$, $\omega_2 = V_{\mathcal{U}}^{1/2} / \mathbb{V}[\hat{\theta}_q]^{1/2}$,

$$\mathcal{S}_n = V_{\mathcal{S}}^{-1/2} \sum_{i=1}^{n} \big(2\tilde{x}_{iq}'\beta - \check{x}_{iq}'\beta\big)\varepsilon_i, \qquad \mathcal{U}_n = V_{\mathcal{U}}^{-1/2} \sum_{i=1}^{n} \sum_{\ell \ne i} C_{i\ell q} \varepsilon_i \varepsilon_\ell,$$

$V_{\mathcal{S}} = \sum_{i=1}^{n}(2\tilde{x}_{iq}'\beta - \check{x}_{iq}'\beta)^2\sigma_i^2$, $V_{\mathcal{U}} = 2\sum_{i=1}^{n}\sum_{\ell\neq i}C_{i\ell q}^2\sigma_i^2\sigma_\ell^2$, and $\mathcal{U}_n$ is uncorrelated of $\mathcal{S}_n$ and $\hat{\mathsf{b}}_q$.

Asymptotic Normality. As in the proof of Theorem 2, we will argue along converging subsequences and therefore move to a subsequence where ω_1 converges. If the limit is zero, then the conclusion of the theorem follows from Lemma B.1 applied to $(\mathbb{V}[v'\hat{\mathsf{b}}_q]^{-1/2}(v'\hat{\mathsf{b}}_q - \mathbb{E}[v'\hat{\mathsf{b}}_q]), \mathcal{U}_n)'$ for $v \in \mathbb{R}^q$ with $v'v = 1$. Thus, we consider the case where the limit of ω_1 is nonzero.

Next, we use Lemma B.1 to show that

$$\left(\mathbb{V}[\hat{\mathsf{b}}_q + u\mathcal{S}_n]^{-1/2}\left(v'\hat{\mathsf{b}}_q - \mathbb{E}\left[v'\hat{\mathsf{b}}_q\right] + u\mathcal{S}_n\right), \mathcal{U}_n\right)' \xrightarrow{d} \mathcal{N}(0, I_2)$$

for any non-random $(v', u)' \in \mathbb{R}^{q+1}$ with $v'v + u^2 = 1$. In the notation of Lemma B.1, we have

$$\dot{w}_i = \mathbb{V}[\hat{\mathsf{b}}_q + u\mathcal{S}_n]^{-1/2}\left(v'\mathsf{w}_{iq} + uV_{\mathcal{S}}^{-1/2}\left(2\tilde{x}_{iq}'\beta - \check{x}_{iq}'\beta\right)\right) \quad \text{and} \quad W_{i\ell} = V_{\mathcal{U}}^{-1/2}C_{i\ell q}.$$

A simple calculation shows that $\mathbb{V}[v'\hat{\mathsf{b}}_q + u\mathcal{S}_n] \geq \min_i \sigma_i^2 \gg 0$, so $\max_i \dot{w}_i^2 = o(1)$ follows from Theorem 3(i), Theorem 3(ii), and ω_1 being bounded away from zero.

Similarly, we have as in the proof of Theorem 2 that

$$\operatorname{trace}\left(C_q^4\right) \leq c\operatorname{trace}\left(B_q^4\right) \leq c\lambda_{q+1}^2\sum_{\ell=q+1}^{r}\lambda_\ell^2 \quad \text{and} \quad V_{\mathcal{U}}^2 \geq \omega_2^{-4}\min_i \sigma_i^8 \operatorname{trace}\left(\tilde{A}^2\right)^2$$

for $C_q = (C_{i\ell q})_{i,\ell}$ and $B_q = (B_{i\ell q})_{i,\ell}$, so Assumptions 1 and 2 yield $\operatorname{trace}(W^4) = o(1)$. *Q.E.D.*

PROOF OF LEMMAS 6 AND 7: See Supplemental Material. *Q.E.D.*

REFERENCES

ABOWD, J. M., R. H. CREECY, F. KRAMARZ et al. (2002): "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Technical report, Center for Economic Studies, US Census Bureau. [1869]

ABOWD, J. M., F. KRAMARZ, AND D. N. MARGOLIS (1999): "High Wage Workers and High Wage Firms," *Econometrica*, 67 (2), 251–333. [1861,1868,1881]

ACHLIOPTAS, D. (2003): "Database-Friendly Random Projections: Johnson-Lindenstrauss With Binary Coins," *Journal of Computer and System Sciences*, 66 (4), 671–687. [1860,1870]

AKRITAS, M. G., AND N. PAPADATOS (2004): "Heteroscedastic One-Way Anova and Lack-of-Fit Tests," *Journal of the American Statistical Association*, 99 (466), 368–382. [1867]

ANATOLYEV, S. (2012): "Inference in Regression Models With Many Regressors," *Journal of Econometrics*, 170 (2), 368–382. [1860,1872,1873]

ANATOLYEV, S., AND M. SØLVSTEN (2020): "Testing Many Restrictions Under Heteroskedasticity," arXiv preprint arXiv:2003.07320. [1873]

ANDREWS, I., AND A. MIKUSHEVA (2016): "A Geometric Approach to Nonlinear Econometric Models," *Econometrica*, 84 (3), 1249–1264. [1861,1877,1878]

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. [1860,1861,1869,1881,1883]

ANGRIST, J., G. IMBENS, AND A. KRUEGER (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14 (1), 57–67. [1863]

ARELLANO, M., AND S. BONHOMME (2011): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *The Review of Economic Studies*, 79 (3), 987–1020. [1867,1868]

BLOOM, N., F. GUVENEN, B. S. SMITH, J. SONG, AND T. VON WACHTER (2018): "The Disappearing Large-Firm Wage Premium," in *AEA Papers and Proceedings*, Vol. 108, 317–322. [1871]

BONHOMME, S., T. LAMADON, AND E. MANRESA (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87 (3), 699–739. [1860,1869]

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 (S1), S13–S70. [1890]

CARD, D., A. R. CARDOSO, AND P. KLINE (2015): "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. [1871,1872]

CARD, D., J. HEINING, AND P. KLINE (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *The Quarterly Journal of Economics*, 128 (3), 967–1015. [1870,1884]

CATTANEO, M. D., M. JANSSON, AND W. K. NEWEY (2018): "Inference in Linear Regression Models With Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113 (523), 1350–1361. [1860,1863,1865,1871,1872]

CHAO, J. C., J. A. HAUSMAN, W. K. NEWEY, N. R. SWANSON, AND T. WOUTERSEN (2014): "Testing Overidentifying Restrictions With Many Instruments and Heteroskedasticity," *Journal of Econometrics*, 178, 15–21. [1860]

CHAO, J. C., N. R. SWANSON, J. A. HAUSMAN, W. K. NEWEY, AND T. WOUTERSEN (2012): "Asymptotic Distribution of Jive in a Heteroskedastic iv Regression With Many Instruments," *Econometric Theory*, 28 (01), 42–86. [1874]

CHATTERJEE, S. (2008): "A New Method of Normal Approximation," *The Annals of Probability*, 36 (4), 1584–1610. [1860]

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, AND D. YAGAN (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project Star," *The Quarterly Journal of Economics*, 126 (4), 1593–1660. [1866,1867]

CHUNG, F. R. (1997): *Spectral Graph Theory. Number 92*. American Mathematical Soc. [1881]

DHAENE, G., AND K. JOCHMANS (2015): "Split-Panel Jackknife Estimation of Fixed-Effect Models," *The Review of Economic Studies*, 82 (3), 991–1030. [1864]

DONALD, S. G., G. W. IMBENS, AND W. K. NEWEY (2003): "Empirical Likelihood Estimation and Consistent Tests With Conditional Moment Restrictions," *Journal of Econometrics*, 117 (1), 55–93. [1873]

DUFOUR, J.-M., AND J. JASIAK (2001): "Finite Sample Limited Information Inference Methods for Structural Equations and Models With Generated Regressors," *International Economic Review*, 42 (3), 815–844. [1877, 1878]

FISHER, R. A. (1925): *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd. [1859,1866]

HAHN, J., AND W. NEWEY (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72 (4), 1295–1319. [1864]

HILDRETH, C., AND J. P. HOUCK (1968): "Some Estimators for a Linear Model With Random Coefficients," *Journal of the American Statistical Association*, 63 (322), 584–595. [1867]

HORN, S. D., R. A. HORN, AND D. B. DUNCAN (1975): "Estimating Heteroscedastic Variances in Linear Models," *Journal of the American Statistical Association*, 70 (350), 380–385. [1860,1863]

JOCHMANS, K., AND M. WEIDNER (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87 (5), 1543–1560. [1860,1880]

JOHNSON, W. B., AND J. LINDENSTRAUSS (1984): "Extensions of Lipschitz Mappings Into a Hilbert Space," *Contemporary Mathematics*, 26 (1), 189–206. [1870]

KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2019a): "Leave-out Estimation of Variance Components," Technical report, National Bureau of Economic Research. [1884,1887,1891]

KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2019b): "LeaveOutTwoWay: A Matlab Package for Leave out Estimation of Variance Components in Two Way Fixed Effects Models," https://github.com/rsaggio87/LeaveOutTwoWay. [1870]

KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): "Supplement to 'Leave-out Estimation of Variance Components'," *Econometrica Supplemental Material*, 88, https://doi.org/10.3982/ECTA16410. [1864,1870]

KUH, E. (1959): "The Validity of Cross-Sectionally Estimated Behavior Equations in Time Series Applications," *Econometrica*, 197–214. [1867]

LACHOWSKA, M., A. MAS, R. D. SAGGIO, AND S. A. WOODBURY (2020): "Do Firm Effects Drift? Evidence From Washington Administrative Data," Technical report, National Bureau of Economic Research. [1884]

LEI, L., P. J. BICKEL, AND N. EL KAROUI (2018): "Asymptotics for High Dimensional Regression m-Estimates: Fixed Design Results," *Probability Theory and Related Fields*, 172 (3-4), 983–1079. [1860]

MACKINNON, J. G., AND H. WHITE (1985): "Some Heteroskedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29 (3), 305–325. [1860, 1863]

MENGER, K. (1927): "Zur allgemeinen kurventheorie," *Fundamenta Mathematicae*, 10 (1), 96–115. [1875]
MOHAR, B. (1989): "Isoperimetric Numbers of Graphs," *Journal of Combinatorial Theory, Series B*, 47 (3), 274–291. [1880]
MOULTON, B. R. (1986): "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 32 (3), 385–397. [1863]
NEWEY, W. K., AND J. R. ROBINS (2018): "Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation," arXiv preprint arXiv:1801.09138. [1860,1862,1874]
PHILLIPS, G. D. A., AND C. HALE (1977): "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems," *International Economic Review*, 219–228. [1863]
POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 1403–1430. [1860,1863]
QUENOUILLE, M. H. (1949): "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society. Series B (Methodological)*, 11 (1), 68–84. [1864]
RAO, C. R. (1970): "Estimation of Heteroscedastic Variances in Linear Models," *Journal of the American Statistical Association*, 65 (329), 161–172. [1863]
RAUDENBUSH, S., AND A. S. BRYK (1986): "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 1–17. [1867,1868]
SARKAR, P., AND P. J. BICKEL (2015): "Role of Normalization in Spectral Clustering for Stochastic Blockmodels," *The Annals of Statistics*, 43 (3), 962–990. [1881]
SCHEFFE, H. (1959): *The Analysis of Variance*. John Wiley & Sons. [1864,1865]
SEARLE, S. R., G. CASELLA, AND C. E. MCCULLOCH (2009): *Variance Components*, Vol. 391. John Wiley & Sons. [1860,1861,1865,1867]
SHERMAN, J., AND W. J. MORRISON (1950): "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *The Annals of Mathematical Statistics*, 21 (1), 124–127. [1862]
SØLVSTEN, M. (2020): "Robust Estimation With Many Instruments," *Journal of Econometrics*, 214 (2), 495–512. [1860,1873,1894]
SWAMY, P. A. (1970): "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 311–323. [1868]
THEIL, H. (1961): Economic forecasts and policy. [1884]
VERDIER, V. (2017): "Estimation and Inference for Linear Models With Two-Way Fixed Effects and Sparsely Matched Data," *Review of Economics and Statistics*. [1863]
WOODBURY, M. A. (1949): "The Stability of out-Input Matrices," Chicago, IL, 9. [1862]
WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT press. [1879]

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 4 June, 2018; final version accepted 26 March, 2020; available online 6 May, 2020.