Leave-out estimation of variance components

Patrick Kline, Raffaele Saggio, Mikkel Sølvsten*

June 4, 2018

Abstract

We propose a framework for unbiased estimation of quadratic forms in the parameters of linear models with many regressors and unrestricted heteroscedasticity. Applications include variance component estimation and tests of linear restrictions in hierarchical and panel models. 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 jackknife bias corrections exhibit first-order biases. The estimator's limiting distribution can be represented by a linear combination of normal and non-central χ^2 random variables. Consistent variance estimators are proposed along with a procedure for constructing uniformly valid confidence intervals. Applying a two-way fixed effects model of wage determination to Italian social security records, we find that ignoring heteroscedasticity substantially biases conclusions regarding the relative contribution of workers, firms, and worker-firm sorting to wage inequality. Monte Carlo exercises corroborate the accuracy of our asymptotic approximations, with clear evidence of non-normality emerging when worker mobility between groups of firms is limited.

Keywords: variance components, heteroscedasticity, fixed effects, leave-out estimation, many regressors, weak identification

*We thank Isaiah Andrews, Bruce Hansen, Whitney Newey, Jack Porter, Andres Santos and seminar participants at UC Berkeley, Harvard, MIT, Princeton, Northwestern, and Wisconsin for helpful comments. The data used in this study was generously provided by the Fondazione Rodolfo De Benedetti. We thank the Berkeley Institute for Research on Labor and Employment for funding support.

As economic datasets 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 (1925)'s 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."[1]

A large experimental literature (Sacerdote, 2001; Graham, 2008; Chetty et al., 2011; Angrist, 2014) employs variants of Fisher's ANOVA approach to infer the degree of variability attributable to peer or classroom effects. Related methods are often used to study heterogeneity across firms, workers, and schools in their responsiveness to exogenous regressors with continuous variation (Raudenbush and Bryk, 1986; Bryk and Raudenbush, 1992; Arellano and Bonhomme, 2011; Graham and Powell, 2012). In labor economics, log-additive models of worker and firm fixed effects are increasingly used to study worker-firm sorting and the dispersion of firm specific pay premia (Abowd et al., 1999; Card et al., 2013, 2018; Song et al., 2017; Sorkin, 2017) and analogous methods have been applied to settings in health (Finkelstein et al., 2016; Silver, 2016) and education (Arcidiacono et al., 2012) economics.

This paper considers 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 are identically distributed draws from a normal distribution. Standard references on this subject (e.g., Searle et al., 2009) suggest diagnostics for heteroscedasticity and non-normality, but offer little guidance regarding estimation and inference when these problems are encountered. Likewise, the econometrics literature on multi-way fixed effects models includes several proposals for the estimation of variance components (Andrews et al., 2008; Jochmans and Weidner, 2016; Bonhomme et al., 2017a,b; Borovičková and Shimer, 2017) but currently provides no approach to conducting inference on these parameters in the plausible setting where heteroscedasticity or non-normality are present.

We begin by proposing a new variance component estimator designed for settings with many regressors and 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) estimates 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 et al., 1989), which allows us to apply recent results on the behavior of second order U-statistics.

We study this 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"

[1] See Cochran (1980) for a discussion of the intellectual development of this early work.

(Cattaneo et al., 2017) or "moderate dimensional" (Lei et al., 2016) asymptotics. We provide verifiable conditions under which the estimator is consistent and show that these conditions are weaker than those required by jackknife bias correction procedures (Quenouille, 1949; Hahn and Newey, 2004; Dhaene and Jochmans, 2015). A series of examples is discussed where the leave-out estimator can be shown to be consistent but such "automatic" bias-correction methods fail due to imbalance in the regressor design.

The large sample distribution of the estimator is derived using a variant of the arguments in Chatterjee (2008) and Sølvsten (2017). In general, this distribution 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. We present conditions under which the limiting distribution simplifies to either a normal or a linear combination of central χ^2 random variables and discuss how these findings can be used to extend existing results on testing linear restrictions (Anatolyev, 2012; Chao et al., 2014; Cattaneo et al., 2017). However, in many settings, neither of these pivotal approximations will be appropriate for conducting inference on variance components.

To construct asymptotically valid confidence intervals in the presence of nuisance parameters, we propose inverting a minimum distance test statistic that utilizes a variance estimator relying on local averages. Critical values are obtained via an application of the procedure of Andrews and Mikusheva (2016). The resulting confidence interval is shown to be uniformly valid 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 et al. (1999) to matched employer employee wage data in a set of Italian provinces. Leave-out estimators find a substantially smaller contribution of firms to wage inequality and much more assortativity in the matching of workers to firms than either the uncorrected plug-in estimator of Abowd et al. (1999) or the homoscedasticity-based correction procedure of Andrews et al. (2008). Monte Carlo exercises utilizing the realized mobility patterns of workers between firms corroborate the accuracy of our asymptotic approximations. 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.

1 Unbiased Estimation of Variance Components

Consider the linear model

$$y_i = x_i'\beta + \varepsilon_i \qquad\qquad (i = 1, \ldots, n)$$

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 non-random symmetric matrix $A \in \mathbb{R}^{k \times k}$ of rank r. We consider either positive semi-definite or non-definite A which, following Searle et al. (2009), correspond to *variance* and *covariance* components respectively. Economic examples where such parameters are of interest are discussed in the next section.

A naive plug-in estimator of θ is the quadratic form $\hat{\theta}_{\text{PI}} = \hat{\beta}' A \hat{\beta}$, where $\hat{\beta} = S_{xx}^{-1} \sum_{i=1}^n x_i y_i$ denotes the 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 $\{\sigma_i^2\}_{i=1}^n$ that takes the form

$$\text{trace}\left(A \mathbb{V}[\hat{\beta}]\right) = \sum_{i=1}^n B_{ii} \sigma_i^2 \quad \text{where} \quad B_{ii} = x_i' S_{xx}^{-1} A S_{xx}^{-1} x_i.$$

As discussed in the next section, this bias can be particularly severe when the dimension of the regressors k is large relative to the sample size.

A bias correction can be motivated by observing that an unbiased estimator of the i-th error variance is

$$\hat{\sigma}_i^2 = y_i \left(y_i - x_i' \hat{\beta}_{-i}\right)$$

where $\hat{\beta}_{-i} = \left(S_{xx} - x_i x_i'\right)^{-1} \sum_{\ell \neq i} x_\ell y_\ell$ denotes the leave-i-out OLS estimator of β. This insight suggests the following bias-corrected estimator of θ:

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

Newey and Robins (2018) observe that "cross-fit" covariances such as $\hat{\sigma}_i^2$ have desirable efficiency properties but we are not aware of existing estimators involving the $\{\hat{\sigma}_i^2\}_{i=1}^n$.

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

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

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

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

An application of the Sherman-Morrison-Woodbury formula (Woodbury, 1949; Sherman and Morrison, 1950) reveals that the representations in (1) and (2) are numerically equivalent:

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

A similar combination of a change of variables argument and a leave-one-out estimator was used by Powell et al. (1989) in the context of weighted average derivatives. The JIVE estimators proposed by Phillips and Hale (1977) and Angrist et al. (1999) also use a leave-one-out estimator, though without the change of variables.[2]

Remark 1. 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. Drineas et al. (2012) provide algorithms to compute these leverages efficiently in large datasets. Spielman and Srivastava (2011) provide analogous methods specialized to the setting where S_{xx} involves a Laplacian matrix, as is often the case in simple two-way fixed effects models (see, e.g., Jochmans and Weidner, 2016, and Appendix B).

Remark 2. In some cases it may be important to allow dependence in the errors in addition to heteroscedasticity. A common case arises when the data are organized into mutually exclusive and independent "clusters" within which the errors may be dependent (Moulton, 1986). The same change of variables argument 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.

1.1 Relation to Existing Approaches

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

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

[2]The object of interest in JIVE estimation is a *ratio* of quadratic forms $\theta_1/\theta_2 = \beta_1'S_{xx}\beta_2/\beta_2'S_{xx}\beta_2$ 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 et al. (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 tend to be biased (see, e.g., Scheffe, 1959, chapter 10, or Appendix C1.1).

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

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

In Section 3 we illustrate that jackknife bias-correction tends to over-correct and produce a first order bias in the opposite direction of the bias in the plug-in estimator. This is analogous to the upward bias in the jackknife estimator of $\mathbb{V}[\hat{\beta}]$ which was derived by Efron and Stein (1981) and shown by Karoui and Purdom (2016) to be of first order importance for inference with many Gaussian regressors.

There are several proposed adaptations of the jackknife to long panels that can decrease bias under stationarity restrictions on the regressors. Letting $t(i) \in \{1, ..., T\}$ denote the time period in which an observation is observed, we can write the panel jackknife of Hahn and Newey (2004) as

$$\hat{\theta}_{\text{PJK}} = T\hat{\theta}_{\text{PI}} - \frac{T-1}{T}\sum_{t=1}^T \hat{\theta}_{\text{PI},-t} \quad \text{where} \quad \hat{\theta}_{\text{PI},-t} = \hat{\beta}_{-t}' A \hat{\beta}_{-t}$$

and $\hat{\beta}_{-t} = \left(\sum_{i:t(i)\neq t} x_i x_i'\right)^{-1} \sum_{i:t(i)\neq t} x_i y_i$ is the OLS estimator that excludes all observations from period t. Dhaene and Jochmans (2015) propose a closely related split panel jackknife

$$\hat{\theta}_{\text{SPJK}} = 2\hat{\theta}_{\text{PI}} - \frac{\hat{\theta}_{\text{PI},1} + \hat{\theta}_{\text{PI},2}}{2} \quad \text{where} \quad \hat{\theta}_{\text{PI},j} = \hat{\beta}_j' A \hat{\beta}_j$$

and $\hat{\beta}_1$ (and $\hat{\beta}_2$) are OLS estimators based on the first half (and the last half) of an even number of time periods. In Section 3, we illustrate how non-stationary regressors or short panels can lead these adaptations of the jackknife to produce first order biases in the opposite direction of the bias in the plug-in estimator.

Remark 3. One might be tempted to estimate θ using the estimators of σ_i^2 employed in Eicker-White style estimators of $\mathbb{V}[\hat{\beta}] = S_{xx}^{-1}\left(\sum_{i=1}^n x_i x_i' \sigma_i^2\right) S_{xx}^{-1}$ (see, e.g., MacKinnon and White (1985) and Davidson and MacKinnon (1993)). Cattaneo et al. (2016) show that the estimation error in $\hat{\beta}$ leads to first order biases in estimators of this type when $k/n \nrightarrow 0$. Their results apply here with

minimal changes since, for a non-random vector v, it follows that:

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

Remark 4. Conversely, one can also use $\{\hat{\sigma}_i^2\}_{i=1}^n$ to construct an unbiased variance estimator $\hat{\mathbb{V}}[\hat{\beta}] = S_{xx}^{-1}\left(\sum_{i=1}^n x_i x_i' \hat{\sigma}_i^2\right) S_{xx}^{-1}$. The variance estimation results in Section 5.2 imply that $\hat{\mathbb{V}}[\hat{\beta}]$ can be used to perform asymptotically valid inference on linear contrasts in settings where existing Eicker-White estimators fail. Specifically, $\hat{\mathbb{V}}[\hat{\beta}]$ leads to valid inference under conditions where the MINQUE estimator of Rao (1970) and the MINQUE-type estimator of Cattaneo et al. (2016) do not exist (see, e.g., Horn et al., 1975; Verdier, 2016).

1.2 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. First, we note that $\hat{\theta}$ is unbiased whenever each of the leave-one-out estimators $\hat{\beta}_{-i}$ exists. This basic requirement is equivalently expressed as $\max_i P_{ii} < 1$ where P_{ii} is the leverage of observation i.

Lemma 1. *If* $\max_i P_{ii} < 1$*, then* $\mathbb{E}[\hat{\theta}] = \theta$.

Next, we show 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 (which is defined by A) and the design matrix S_{xx}.

Let $\lambda_1, \ldots, \lambda_r$ denote the non-zero 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 $D = \text{diag}(\lambda_1, \ldots, \lambda_r)$. With these definitions we have

$$\hat{\beta}' A \hat{\beta} = \sum_{\ell=1}^{r} \lambda_\ell \hat{b}_\ell^2 \quad \text{where} \quad \hat{b} = (\hat{b}_1, \ldots, \hat{b}_r)' = Q' S_{xx}^{1/2} \hat{\beta}.$$

The r-dimensional random vector $\hat{b}$ and the eigenvalues $\lambda_1, \ldots, \lambda_r$ are central to both the finite sample distribution provided below in Lemma 2 and the asymptotic properties of $\hat{\theta}$ as studied in Section 4. Each eigenvalue of $\tilde{A}$ can be thought of as measuring 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 $S_{xx}^{-1/2}$). As discussed in Section 4, when a few of these eigenvalues are large relative to the others, a form of weak identification can arise.

Lemma 2. *If* $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$*, then*

$$\theta^* = \sum_{\ell=1}^{r} \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell] \right) \quad \textit{and} \quad \hat{b} \sim \mathcal{N}\left(b, \mathbb{V}[\hat{b}] \right)$$

where $b = Q' S_{xx}^{1/2} \beta$.

The distribution of θ^* is a sum of r potentially dependent non-central χ^2 random variables with non-centralities $b = (b_1, \ldots, b_r)'$. In the special case of homoscedasticity ($\sigma_i^2 = \sigma^2$) and no signal ($b = 0$) we have that $\hat{b} \sim \mathcal{N}\left(0, \sigma^2 I_r\right)$, 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 squared eigenvalues converge 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.

1.3 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:

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

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 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 $\left\{\varepsilon_{i,n} : 1 \leq i \leq n, n \geq 1\right\}$. This *conditional* perspective is common in the statistics literatures on ANOVA (Scheffe, 1959; Searle et al., 2009) and high-dimensional models (Lei et al., 2016), and allows us to be agnostic about the potential dependency among the $\{x_{i,n}\}_{i=1}^n$. Following standard practice we drop the n subscript in what follows. All limits are taken as n goes to infinity unless otherwise noted.

Our analysis makes heavy use of the following assumptions.

Assumption 1. *(a)* $\max_i \left(\mathbb{E}[\varepsilon_i^4] + \sigma_i^{-2} \right) = O(1)$*, (b) there exist a* $c < 1$ *such that* $\max_i P_{ii} \leq c$ *for all* n*, and (c)* $\max_i (x_i' \beta)^2 = O(1)$.

Part (a) of this condition limits the thickness of the tails in the error distribution, as is typically required for OLS estimation (see, e.g., Cattaneo et al., 2017, page 10). The bounds on $(x_i'\beta)^2$ and P_{ii} imply that $\hat{\sigma}_i^2$ has bounded variance. Assumption 1(c) is a technical condition that can be relaxed to allow for $\max_i(x_i'\beta)^2$ to be unbounded as the sample size grows as discussed further in Section 6. From (b) it follows that $\frac{k}{n} \leq c < 1$ for all n.

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

Lemma 3. *1. If A is positive semi-definite, (i) $\theta = O(1)$,*

$$(ii)\ \textit{trace}(\tilde{A}^2) = \sum_{\ell=1}^{r} \lambda_\ell^2 = o(1),$$

and Assumption 1 holds, then $\hat{\theta} - \theta \xrightarrow{p} 0$.

2. If A is non-definite then write $A = A_1'A_2$ for some A_1, A_2. If $\Theta_\ell = \beta' A_\ell' A_\ell \beta$ satisfies (i) and (ii) for $\ell = 1, 2$, then $\hat{\theta} - \theta \xrightarrow{p} 0$.

The first part of the Lemma establishes consistency of variance components given boundedness of θ and a joint condition on the design matrix S_{xx} and the matrix A. In several of the examples discussed in the next section, $\text{trace}(\tilde{A}^2)$ is of order r/n^2 which automatically satisfies (ii). A more extensive discussion of primitive conditions that yield (ii) is provided in Section 6. Consistency of covariance components follows from consistency of variance components that dominate them via the Cauchy-Schwarz inequality, i.e., $\theta^2 \leq \Theta_1\Theta_2$.

2 Examples

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

Example 1 (Coefficient of determination)**.**

Sewall Wright (1921) proposed measuring the explanatory power of a linear model using the coefficient of determination. When x_i includes an intercept, the object of interest and its corresponding plug-in estimator can be written

$$R^2 = \frac{\beta' A \beta}{\beta' A \beta + \frac{1}{n}\sum_{i=1}^n \sigma_i^2} = \frac{\sigma_{X\beta}^2}{\sigma_y^2} \quad \text{and} \quad \hat{R}^2_{\text{PI}} = \frac{\hat{\beta}' A \hat{\beta}}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\hat{\sigma}^2_{X\beta,\text{PI}}}{\hat{\sigma}_y^2}$$

where

$$A = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})', \quad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

Theil (1961) noted that the plug-in estimator of $\sigma^2_{X\beta}$ is biased and proposed an adjusted R^2 measure that utilizes the homoscedasticity-only estimator in (3)

$$\hat{R}^2_{\text{adj}} = \frac{\hat{\sigma}^2_{X\beta,\text{HO}}}{\hat{\sigma}^2_y} = \frac{\hat{\beta}' A \hat{\beta} - \frac{k-1}{n}\hat{\sigma}^2_{\text{HO}}}{\hat{\sigma}^2_y} \quad \text{where} \quad \sum_{i=1}^n B_{ii} = \frac{k-1}{n}.$$

A rearrangement gives the familiar representation $\frac{1-\hat{R}^2_{\text{adj}}}{1-\hat{R}^2_{\text{PI}}} = \frac{n-1}{n-k}$ which highlights that the adjusted estimator of R^2 relates to the unadjusted one through a degrees-of-freedom correction.

The leave-out estimator of $\sigma^2_{X\beta}$ allows for unrestricted heteroscedasticity and can be found by noting that $\tilde{x}_i = AS^{-1}_{xx}x_i = \frac{1}{n}(x_i - \bar{x})$, which yields

$$\hat{R}^2 = \frac{\hat{\sigma}^2_{X\beta}}{\hat{\sigma}^2_y} \quad \text{where} \quad \hat{\sigma}^2_{X\beta} = \frac{1}{n}\sum_{i=1}^n y_i (x_i - \bar{x})' \hat{\beta}_{-i}.$$

In general, this estimator does not have an interpretation in terms of degrees-of-freedom corrections. Instead, the explanatory power of the linear model is assessed using the empirical covariance between leave-one-out predictions $(x_i - \bar{x})'\hat{\beta}_{-i}$ and the left out observation y_i.

Example 2 (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 dataset comprised of observations on N groups with T_g observations in the g-th group. The "analysis of covariance" model posits that outcomes can be written

$$y_{gt} = \alpha_g + x'_{gt}\delta + \varepsilon_{gt} \qquad (g = 1, \ldots, N,\ t = 1, \ldots, T_g \geq 2),$$

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

A recent application comes from Chetty et al. (2011) who study 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:

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

where $\bar{\alpha} = \frac{1}{n}\sum_{g=1}^N T_g \alpha_g$ gives the (enrollment-weighted) mean classroom effect. This model and object of interest can be brought in to the notation of the preceding section ($y_i = x'_i\beta + \varepsilon_i$ and $\sigma^2_\alpha = \beta' A \beta$) if for each (g,t) we let $i = i(g,t)$ where $i(\cdot,\cdot)$ is bijective with inverse denoted $(g(\cdot), t(\cdot))$,

$y_i = y_{gt}$, $\varepsilon_i = \varepsilon_{gt}$,

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

and

$$A = \begin{bmatrix} A_{dd} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{where} \quad A_{dd} = \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})(d_i - \bar{d})', \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i.$$

Chetty et al. (2011) estimate σ_α^2 using a random effects ANOVA estimator (see e.g., Searle et al., 2009) which is of the homoscedasticity-only type given in (3). As discussed in Section 1 and Appendix C1.1, this estimator is in general first order biased when the errors are heteroscedastic and group sizes are unbalanced.

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:

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

where $\hat{\alpha}_g = \frac{1}{T_g}\sum_{t=1}^{T_g} y_{gt}$, and $\hat{\bar{\alpha}} = \frac{1}{n}\sum_{g=1}^{N} T_g \hat{\alpha}_g$. This representation shows that if the model consists of group specific intercepts only, 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.

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

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

The generalized regressor $\tilde{d}_i$ can be described as follows: if there are no common regressors then $\tilde{d}_i = \frac{1}{n}(d_i - \bar{d})$, which is analogous to Example 1. If the model includes common regressors then $\tilde{d}_i = \frac{1}{n}\left((d_i - \bar{d}) - \hat{\Gamma}'(x_{g(i)t(i)} - \bar{x}_{g(i)})\right)$ where $\bar{x}_g = \frac{1}{T_g}\sum_{t=1}^{T_g} x_{gt}$ and $\hat{\Gamma}$ is the coefficient vector from an instrumental variables (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, because d_i is uncorrelated with $x_{g(i)t(i)} - \bar{x}_{g(i)}$, the covariance between d_i and $\tilde{d}_i$ is A_{dd} . This ensures that the empirical covariance between $y_i = d_i'\alpha + x_{g(i)t(i)}'\delta + \varepsilon_i$ and the generalized prediction $\tilde{d}_i'\hat{\alpha}_{-i}$ is an unbiased estimator of σ_α^2.

Example 3 (Random coefficients)**.**

Group memberships are often modeled as influencing slopes in addition to intercepts (Kuh, 1959;

Hildreth and Houck, 1968; Bryk and Raudenbush, 1992; Arellano and Bonhomme, 2011; Graham and Powell, 2012; Graham et al., 2016). Consider the following "random coefficient" model:

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

with $n = \sum_{g=1}^{N} T_g$.

An influential example comes from Raudenbush and Bryk (1986), who model student mathematics scores as a "hierarchical" linear function of socioeconomic status (SES) with school-specific intercepts ($\alpha_g \in \mathbb{R}$) and slopes ($\gamma_g \in \mathbb{R}$). The student-weighted variance of slopes can be written:

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

where $\bar{\gamma} = \frac{1}{n}\sum_{g=1}^{N} T_g \gamma_g$. In the notation of the preceding section we can write $y_i = x_i'\beta + \varepsilon_i$ and $\sigma_\gamma^2 = \beta' A \beta$ where

$$x_i = (d_i', d_i' z_{gt})', \qquad \beta = (\alpha', \gamma')', \quad \gamma = (\gamma_1, \ldots, \gamma_N)', \qquad A = \begin{bmatrix} 0 & 0 \\ 0 & A_{dd} \end{bmatrix}$$

for y_i, ε_i, d_i, A_{dd}, and α as in the preceding example.

Raudenbush and Bryk (1986) use a maximum likelihood estimator of σ_γ^2 predicated upon normality and homoscedastic errors. Swamy (1970) considers 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}. By contrast, the leave-out estimator is unbiased under arbitrary patterns of heteroscedasticity.

The leave-out estimator can be represented in terms of the empirical covariance

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

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 makes $\tilde{d}_i \tilde{z}_i$ uncorrelated with d_i and 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_{dd}. This implies that the empirical covariance between $y_i = d_i'\alpha + z_{g(i)t(i)} d_i'\gamma + \varepsilon_i$ and the generalized prediction $\tilde{z}_i \tilde{d}_i' \hat{\gamma}_{-i}$ is an unbiased estimator of σ_γ^2.

Example 4 (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 et al. (1999) (henceforth AKM) who propose 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:

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + x'_{gt}\delta + \varepsilon_{gt} \qquad (g = 1, \ldots, N,\ t = 1, \ldots, T_g \geq 2) \tag{5}$$

where the function $j(\cdot,\cdot) : \{1, \ldots, N\} \times \{1, \ldots, \max_g T_g\} \to \{0, \ldots, J\}$ allocates each of $n = \sum_{g=1}^{N} T_g$ worker-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 (see, e.g., Card et al., 2013; Song et al., 2017). AKM summarize the firm contribution to wage inequality via the following two parameters:

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

where $\bar{\psi} = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}\psi_{j(g,t)}$. The variance component σ^2_ψ measures the 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 and the corresponding objects of interest in the notation of the preceding section ($y_i = x_i'\beta + \varepsilon_i$, $\sigma^2_\psi = \beta' A_\psi \beta$, and $\sigma_{\alpha,\psi} = \beta' A_{\alpha,\psi}\beta$), let

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

for y_i, ε_i, and d_i as in the preceding examples,

$$f_i = (\mathbf{1}_{\{j(g,t)=1\}}, \ldots, \mathbf{1}_{\{j(g,t)=J\}})',$$

$$A_\psi = \begin{bmatrix} 0 & 0 & 0 \\ 0 & A_{ff} & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{where} \quad A_{ff} = \frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})(f_i - \bar{f})', \quad \bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i,$$

and

$$A_{\alpha,\psi} = \frac{1}{2}\begin{bmatrix} 0 & A_{df} & 0 \\ A'_{df} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{where} \quad A_{df} = \frac{1}{n}\sum_{i=1}^{n} d_i(f_i - \bar{f})'.$$

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. (1999, 2002) note, 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.

AKM estimate σ^2_ψ 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, in general, biased. Andrews et al. (2008) propose the "homoscedasticity-only" estimators of (3). These estimators are unbiased when the errors ε_i are independent and have common variance. Our leave-out estimator, which avoids the homoscedasticity requirement on the errors, takes the form

$$\hat{\sigma}^2_\psi = \sum_{i=1}^{n} y_i x_i' S_{xx}^{-1} A_\psi \hat{\beta}_{-i}, \qquad \hat{\sigma}_{\alpha,\psi} = \sum_{i=1}^{n} y_i x_i' S_{xx}^{-1} A_{\alpha,\psi} \hat{\beta}_{-i}. \tag{6}$$

A simpler representation of $\hat{\sigma}^2_\psi$ 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

$$\Delta y_g = \Delta f_g' \psi + \Delta \varepsilon_g \qquad\qquad (g = 1, \ldots, N) \tag{7}$$

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 estimator $\hat{\sigma}^2_\psi$ equals the leave-out estimator of σ^2_ψ applied to this model:

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

$S_{\Delta f \Delta f}$ and $\hat{\psi}_{-g}$ correspond respectively to S_{xx} and $\hat{\beta}_{-i}$ in the above first differenced model. This equivalence reveals that $\hat{\sigma}^2_\psi$ is not only unbiased 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 on observations with $\Delta f_g \neq 0$ (i.e., firm "movers").

3 Comparison to Jackknife Estimators

This section compares the leave-out estimator $\hat{\theta}$ to estimators predicated on jackknife bias corrections. We start by introducing some of the high-level assumptions that are typically used to motivate jackknife estimators. We then consider some variants of Examples 2 and 3 where these high-level conditions fail to hold and establish that the jackknife estimators have first order biases while the leave-out estimator retains consistency.

3.1 High-level Conditions

Jackknife bias corrections are typically motivated by the high-level assumption that the bias of a plug-in estimator $\hat{\theta}_{\text{PI}}$ shrinks with the sample size in a known way and that the bias of $\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{\text{PI},-i}$ depends on sample size in an identical way, i.e.,

$$\mathbb{E}[\hat{\theta}_{\text{PI}}] = \theta + \frac{\text{D}_1}{n} + \frac{\text{D}_2}{n^2}, \quad \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{\text{PI},-i}\right] = \theta + \frac{\text{D}_1}{n-1} + \frac{\text{D}_2}{(n-1)^2} \quad \text{for some } \text{D}_1, \text{D}_2. \tag{8}$$

Under (8), the jackknife estimator $\hat{\theta}_{\text{JK}} = n\hat{\theta}_{\text{PI}} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{\text{PI},-i}$ has a bias of $-\frac{\text{D}_2}{n(n-1)}$.

For some long panel settings the bias in $\hat{\theta}_{\text{PI}}$ is shrinking in the number of time periods T such that

$$\mathbb{E}[\hat{\theta}_{\text{PI}}] = \theta + \frac{\dot{\text{D}}_1}{T} + \frac{\dot{\text{D}}_2}{T^2} \quad \text{for some } \dot{\text{D}}_1, \dot{\text{D}}_2.$$

In such settings, it may be that the biases of $\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_{\text{PI},-t}$ and $\frac{1}{2}(\hat{\theta}_{\text{PI},1} + \hat{\theta}_{\text{PI},2})$ depend on T in an identical way, i.e.,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_{\text{PI},-t}\right] = \theta + \frac{\dot{\text{D}}_1}{T-1} + \frac{\dot{\text{D}}_2}{(T-1)^2} \quad \text{and} \quad \mathbb{E}\left[\frac{1}{2}(\hat{\theta}_{\text{PI},1} + \hat{\theta}_{\text{PI},2})\right] = \theta + \frac{2\dot{\text{D}}_1}{T} + \frac{4\dot{\text{D}}_2}{T^2}.$$

From here it follows that the panel jackknife estimator $\hat{\theta}_{\text{PJK}} = T\hat{\theta}_{\text{PI}} - \frac{T-1}{T}\sum_{t=1}^{T}\hat{\theta}_{\text{PI},-t}$ has a bias of $-\frac{\dot{\text{D}}_2}{T(T-1)}$ and that the split panel jackknife estimator $\hat{\theta}_{\text{SPJK}} = 2\hat{\theta}_{\text{PI}} - \frac{1}{2}(\hat{\theta}_{\text{PI},1} + \hat{\theta}_{\text{PI},2})$ has a bias of $-\frac{2\dot{\text{D}}_2}{T^2}$, both of which shrink faster to zero than $\frac{\dot{\text{D}}_1}{T}$ if $T \to \infty$. Typical sufficient conditions for bias-representations of this kind to hold (to second order) are that (a) $T \to \infty$, (b) the design is stationary over time, and (c) that $\hat{\theta}_{\text{PI}}$ is asymptotically linear (see, e.g., Hahn and Newey, 2004; Dhaene and Jochmans, 2015). Below we illustrate that jackknife corrections can be inconsistent in Examples 2 and 3 when (a) and/or (b) do not hold. Finally we note that $\hat{\theta}_{\text{PI}}$ (a bilinear function) need not be asymptotically linear as is evident from the non-normal asymptotic distribution of $\hat{\theta}$ derived in Theorem 1 of the next section.

3.2 Examples of Jackknife Failure

Example 2 (Special case)**.** Consider the model

$$y_{gt} = \alpha_g + \varepsilon_{gt} \qquad (g = 1, \ldots, N,\ t = 1, \ldots, T \geq 2),$$

where $\sigma_{gt}^2 = \sigma^2$ and suppose the parameter of interest is $\theta = \frac{1}{N}\sum_{g=1}^N \alpha_g^2$. For T even, we have the following bias calculations:

$$\mathbb{E}[\hat{\theta}_{\mathrm{PI}}] = \theta + \frac{\sigma^2}{T}, \qquad \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \hat{\theta}_{\mathrm{PI},-i}\right] = \theta + \frac{\sigma^2}{T} + \frac{\sigma^2}{n(T-1)},$$
$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \hat{\theta}_{\mathrm{PI},-t}\right] = \theta + \frac{\sigma^2}{T-1}, \qquad \mathbb{E}\left[\frac{1}{2}(\hat{\theta}_{\mathrm{PI},1} + \hat{\theta}_{\mathrm{PI},2})\right] = \theta + \frac{2\sigma^2}{T}.$$

The jackknife estimator $\hat{\theta}_{JK}$ has a first order bias of $-\frac{\sigma^2}{T(T-1)}$, which when $T = 2$ is as large as that of $\hat{\theta}_{\mathrm{PI}}$ but of opposite sign. By contrast, both of the panel jackknife estimators, $\hat{\theta}_{PJK}$ and the leave-out estimator are exactly unbiased and consistent as $n \to \infty$ when T is fixed.

This example shows that the jackknife estimator can fail when applied to a setting where the number of regressors is large relative to sample size. Here the number of regressors is N and the sample size is NT, yielding a ratio of $1/T$ and we see that $1/T \to 0$ is necessary for consistency of $\hat{\theta}_{\mathrm{JK}}$. While the panel jackknife corrections appear to handle the presence of many regressors, this property disappears in the next example which adds the "random coefficients" of Example 3.

Example 3 (Special case)**.** Consider the model

$$y_{gt} = \alpha_g + x_{gt}\delta_g + \varepsilon_{gt} \qquad (g = 1, \ldots, N,\ t = 1, \ldots, T \geq 3)$$

where $\sigma_{gt}^2 = \sigma^2$ and $\theta = \frac{1}{N}\sum_{g=1}^N \delta_g^2$.

An analytically convenient example arises when the regressor design is "balanced" across groups as follows:

$$(x_{g1}, x_{g2}, \ldots, x_{gT}) = (x_1, x_2, \ldots, x_T),$$

where x_1, x_2, x_3 take distinct values and $\sum_{t=1}^T x_t = 0$. The leave-out estimator is unbiased and consistent for any $T \geq 3$, whereas for even $T \geq 4$ we have the following bias calculations:

$$\mathbb{E}[\hat{\theta}_{\mathrm{PI}}] = \theta + \frac{\sigma^2}{\sum_{t=1}^T x_t^2},$$
$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \hat{\theta}_{\mathrm{PI},-t}\right] = \theta + \frac{\sigma^2}{T}\sum_{t=1}^T \frac{1}{\sum_{s\neq t}(x_s - \bar{x}_{-t})^2},$$
$$\mathbb{E}\left[\frac{1}{2}(\hat{\theta}_{\mathrm{PI},1} + \hat{\theta}_{\mathrm{PI},2})\right] = \theta + \frac{\sigma^2}{2\sum_{t=1}^{T/2}(x_t - \bar{x}_1)^2} + \frac{\sigma^2}{2\sum_{t=T/2+1}^{T}(x_t - \bar{x}_2)^2},$$

where $\bar{x}_{-t} = \frac{1}{T-1}\sum_{s\neq t} x_s$, $\bar{x}_1 = \frac{2}{T}\sum_{t=1}^{T/2} x_t$, and $\bar{x}_2 = \frac{2}{T}\sum_{t=T/2+1}^{T} x_t$.

The calculations above reveal that non-stationarity in either the level or variability of x_t over time can lead to a negative bias in panel jackknife approaches, e.g.,

$$\mathbb{E}\left[\hat{\theta}_{\text{SPJK}}\right] - \theta \leq \frac{2\sigma^2}{\sum_{t=1}^{T} x_t^2} - \frac{\sigma^2}{2\sum_{t=1}^{T/2} x_t^2} - \frac{\sigma^2}{2\sum_{t=T/2+1}^{T} x_t^2} \leq 0$$

where the first inequality is strict if $\bar{x}_1 \neq \bar{x}_2$ and the second if $\sum_{t=1}^{T/2} x_t^2 \neq \sum_{t=T/2+1}^{T} x_t^2$. In fact, the following example

$$(x_1, x_2, \ldots, x_T) = (-1, 2, 0, \ldots, 0, -1)$$

renders the panel jackknife corrections inconsistent for small or large T:

$$\mathbb{E}[\hat{\theta}_{\text{PJK}}] = \theta - \frac{7/5}{6}\sigma^2 + O\left(\frac{1}{T}\right) \quad \text{and} \quad \mathbb{E}[\hat{\theta}_{\text{SPJK}}] = \theta - \frac{8/5}{6}\sigma^2 + O\left(\frac{1}{T}\right).$$

Inconsistency results here from biases of first order that are negative and larger in magnitude than the original bias of $\hat{\theta}_{\text{PI}}$ (which is $\frac{\sigma^2}{6}$). Exact bias formulas are given in Appendix C3.

4 Distribution theory

In this section, we develop asymptotic theory intended to approximate the finite sample behavior of $\hat{\theta}$ in a wide array of settings. Section 1.2 showed that the finite sample distribution of the infeasible estimator θ^* under normality of the errors is a sum of r non-central χ^2 random variables weighted by the eigenvalues $\lambda_1, \ldots, \lambda_r$ of $\tilde{A}$, i.e., for $\theta^* = \hat{\beta}' A \hat{\beta} - \sum_{i=1}^{n} B_{ii}\sigma_i^2$ and $B_{ii} = x_i' S_{xx}^{-1} A S_{xx}^{-1} x_i$ we have

$$\theta^* = \sum_{\ell=1}^{r} \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell]\right) \quad \text{and} \quad \hat{b} \sim \mathcal{N}\left(b, \mathbb{V}[\hat{b}]\right)$$

where $b = Q' S_{xx}^{1/2} \beta$. This distribution is centered at θ, but its shape depends on the r-dimensional nuisance parameter b, which complicates using this result for inference. When r is small, a potential approach is to base inference on a minimum distance statistic for $\hat{b}$. In general, this approach need not have any optimality properties as $\theta = \sum_{\ell=1}^{r} \lambda_\ell b_\ell^2$ is a non-invertible function of b, but it can be shown to be asymptotically valid when the estimator of $\mathbb{V}[\hat{b}]$ utilizes $\{\hat{\sigma}_i^2\}_{i=1}^n$. In many applications, however, r will be large and computation of $\hat{b}$ will become intractable because it involves *all* the eigenvectors of $\tilde{A}$. We therefore provide asymptotic approximations to the distribution of $\hat{\theta}$ that serve both to relax the normality assumption on the errors and to motivate an inference procedure based on a minimum distance statistic for a vector of substantially lower dimension than $\hat{b}$.

In Proposition 1 we show that the finite sample distribution of θ^* provides a good approximation to the asymptotic distribution of $\hat{\theta}$ when r is small. Proposition 2 establishes that when r is large, and the largest squared eigenvalue λ_1^2 is small relative to the sum of squared eigenvalues $\sum_{\ell=1}^r \lambda_\ell^2$, the asymptotic distribution of $\hat{\theta}$ simplifies to that of a normal random variable. Approximations of these two kinds are common in the literature on hypothesis testing (see, e.g., Andrews, 1988; Anatolyev, 2012; Chao et al., 2014), but we are not aware of existing theorems that contain our results as special cases.

Propositions 1 and 2 are important in their own right as the objects of interest in Examples 1 and 2 are covered by these results. However, these results also serve to motivate Theorem 1, which covers the case where r is large and *some* of the squared eigenvalues are large relative to their sum. In this case the resulting asymptotic distribution is a linear combination of normal and non-central χ^2 random variables. This added generality is necessary to accommodate Examples 3 and 4 and we are not aware of existing results that provide approximations of this type.

4.1 The low rank case

The following Proposition characterizes the asymptotic distribution of $\hat{\theta}$ when r is small. The result relies on the observation that $\hat{b}$ is a weighted sum,

$$\hat{b} = \sum_{i=1}^n w_i y_i \quad \text{where} \quad w_i = (w_{i1}, \ldots, w_{ir})' = Q' S_{xx}^{-1/2} x_i,$$

which is asymptotically normal when no observation is too influential, i.e., when $\max_i w_i' w_i = o(1)$. 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 $w_{i\ell}^2 = \frac{1}{n}$.

Proposition 1. *If Assumption* 1 *holds,* $\max_i w_i' w_i = o(1)$*, and* r *is fixed, then*

$$\hat{\theta} = \sum_{\ell=1}^r \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell] \right) + o_p(\mathbb{V}[\hat{\theta}]^{1/2}) \quad \textit{and} \quad \mathbb{V}[\hat{b}]^{-1/2} (\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, I_r)$$

where $b = Q' S_{xx}^{1/2} \beta$*, and* $\mathbb{V}[\hat{b}] = \sum_{i=1}^n w_i w_i' \sigma_i^2$*.*

Here, the limit distribution of $\hat{\theta}$ is first order equivalent to that derived for the infeasible estimator θ^* when $\hat{\beta}$ is normally distributed. However, Proposition 1 allows $\hat{\beta}$ to include statistics estimated from as few as two observations, so $\hat{\beta}$ need not behave as a normally distributed vector in large samples. Rather, the assumptions imply that the r-dimensional vector $\hat{b}$ is approximately normal and that the estimated bias correction $\sum_{i=1}^n B_{ii} \hat{\sigma}_i^2$ has a second order effect on the variability of $\hat{\theta}$.

Depending upon the nature of the target parameter θ, the condition $\max_i w_i'w_i = o(1)$ may directly constrain the limiting behavior of $\hat{\beta}$. For example, if A is such that θ corresponds to the square of the first element of β, this condition requires that the first element of $\hat{\beta}$ (though not the other elements) be asymptotically normally distributed. By contrast, if θ corresponds to the square of an average of several elements in β, then all that is needed is for the average of these elements to be asymptotically normal.

Since $\frac{1}{n}\sum_{i=1}^n w_i'w_i = \frac{r}{n}$ we have that the Lindeberg condition $\max_i w_i'w_i = o(1)$ 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). On the other hand, this observation also makes it clear that Proposition 1 does not apply to settings where r is proportional to n as $\max_i w_i'w_i \geq \frac{r}{n}$.

Remark 5. Proposition 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. In such a setting a natural choice of A 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 an equally 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. For a recent contribution to this literature, see Anatolyev (2012) who allows for many regressors but considers the special case of homoscedastic errors.

4.2 The high rank case

For the next two results, let $\check{x}_i = \sum_{\ell=1}^n M_{i\ell}\frac{B_{\ell\ell}}{1-P_{\ell\ell}}x_\ell$ where $M_{i\ell} = \mathbf{1}\{i=\ell\} - x_iS_{xx}^{-1}x_\ell$. 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 could 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 the following Proposition, this estimation error can resurface in the asymptotic distribution.

Proposition 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*

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

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

One can think of the eigenvalue ratio in (ii) as the inverse effective rank of $\tilde{A}$: when all the eigenvalues are equal $\frac{\lambda_1^2}{\sum_{\ell=1}^r\lambda_\ell^2} = \frac{1}{r}$. Moreover, (ii) is a Lindeberg condition which ensures that the weighted sum $\sum_{\ell=1}^r\lambda_\ell\hat{b}_\ell^2$ is not dominated by any of the random variables $\hat{b}_1,\ldots,\hat{b}_r$. However, the random variables $\hat{b}_1,\ldots,\hat{b}_r$ are not necessarily independent, which renders the classical Lindeberg

central limit theorem inapplicable. Instead the proof of Proposition 2 relies on a variation of Stein's method developed in Sølvsten (2017) and a representation of $\hat{\theta}$ as a second order U-statistic, i.e.,

$$\hat{\theta} = \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} y_i y_\ell \tag{9}$$

where $C_{i\ell} = B_{i\ell} - 2^{-1} M_{i\ell} \left(M_{ii}^{-1} B_{ii} + M_{\ell\ell}^{-1} B_{\ell\ell} \right)$ 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 Proposition 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 (i) will be satisfied whenever $\max_i \left((\tilde{x}_i'\beta)^2 + (\check{x}_i'\beta)^2 \right)$ is not too large compared to $\sum_{i=1}^{n} (\tilde{x}_i'\beta)^2 + (\check{x}_i'\beta)^2$. As in Proposition 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 $\beta = 0$.

Remark 6. A natural application of Proposition 2 is to tests of specification that can be formulated in terms of 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 $A = \frac{1}{r} R' (R S_{xx}^{-1} R')^{-1} R$ implies $\mathbb{V}[\hat{\theta}]^{-1/2} \hat{\theta} \xrightarrow{d} \mathcal{N}(0, 1)$ since all the non-zero eigenvalues of $\tilde{A}$ are equal to $\frac{1}{r}$. The existing literature allows for either heteroscedastic errors and moderately few regressors (Donald et al., 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 Section 5, this result enables tests with heteroscedastic errors and many regressors.

Remark 7. Proposition 2 extends some common results in the literature on many and many weak instruments (see, e.g., Chao et al., 2012) where the estimators are asymptotically equivalent to bilinear forms. The structure of that setting is such that $\tilde{A} = I_r/r$ and $r \to \infty$, in which case Proposition 2(ii) is automatically satisfied and therefore does not feature prominently in that literature.

4.3 The general case

We turn now to our main theorem which covers the case where some of the squared eigenvalues $\lambda_1^2, \ldots, \lambda_r^2$ are large relative to their sum $\sum_{\ell=1}^r \lambda_\ell^2$. To motivate this condition, 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 $S_{xx}^{-1/2}$). From Lemma 3, $\text{trace}(\tilde{A}^2) = \sum_{\ell=1}^r \lambda_\ell^2$ governs the total variability in $\hat{\theta}$. Therefore, Theorem 1 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 $\hat{\theta}$. We discuss in Section 6 when this state of affairs can arise.

Assumption 2. *There exist a $c > 0$ and a known and fixed $q \in \{1, \ldots, r-1\}$ 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 $\hat{\theta}$. The assumption that q is known is motivated by our application and the discussion of Examples 1-4 in Section 6. In Section 5.3 we offer some guidance on choosing q in settings where it is unknown.

Given knowledge of q, we can split $\hat{\theta}$ into a known function of $\hat{\mathsf{b}}_q$ and $\hat{\theta}_q$ where

$$\hat{\mathsf{b}}_q = (\hat{b}_1, \ldots, \hat{b}_q)' = \sum_{i=1}^n \mathsf{w}_{iq} y_i, \qquad \mathsf{w}_{iq} = (w_{i1}, \ldots, w_{iq})',$$

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

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 Proposition 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 1. It is possible to provide a theorem that simultaneously covers Proposition 1 ($q = r$, r fixed) and Proposition 2 ($q = 0$, $r \to \infty$), but to avoid dealing with settings where $\hat{\mathsf{b}}_q$ is an empty vector or $\hat{\theta}_q$ is identically zero we exclude these cases below.

Theorem 1. *If* $\max_i \mathsf{w}_{iq}' \mathsf{w}_{iq} = o(1)$, $\mathbb{V}[\hat{\theta}_q]^{-1} \max_i \left((\tilde{x}_{iq}' \beta)^2 + (\check{x}_{iq}' \beta)^2 \right) = o(1)$, *and Assumptions* 1

and 2 *hold, then*

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

and

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

Theorem 1 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 $\mathbb{E}[\hat{\mathsf{b}}_q] = (b_1, \ldots, b_q)'$ 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.

5 Inference

In this section, we 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 1 implies are jointly normally distributed. To avoid the conservatism associated with standard projection methods (e.g., Dufour and Jasiak, 2001), we seek to 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 (e.g., the problem of subvector inference), θ is a nonlinear function of $\mathbb{E}[\hat{\mathsf{b}}_q]$. To accomodate this complication, we use a critical value proposed by Andrews and Mikusheva (2016) that depends on the curvature of the problem.

When $q = 0$, this procedure simplifies to a standard two-sided confidence interval based on $\hat{\theta}$ and asymptotic normality. If $q = 1$, the resulting confidence interval has a closed form solution, and for $q > 1$, inference relies on solving two quadratic optimization problems that involve $q + 1$ unknowns. Here we focus on the cases of $q = 0$ and $q = 1$ and relegate the full description of the case where $q > 1$ to Appendix C5.2.

5.1 Confidence Interval

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 confi-

dence, $1-\alpha$, we consider the interval

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

where

$$\mathsf{B}_q = \left\{ (\mathsf{b}_q', \theta_q)' \in \mathbb{R}^{q+1} : \begin{pmatrix} \hat{\mathsf{b}}_q - \mathsf{b}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix}' \hat{\Sigma}_q^{-1} \begin{pmatrix} \hat{\mathsf{b}}_q - \mathsf{b}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix} \leq z_{\hat{\kappa}}^2 \right\}$$

and $\hat{\Sigma}_q = \hat{\mathbb{V}}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)']$ is an estimator of $\Sigma_q = \mathbb{V}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)']$ given in the next subsection.

The critical value function, z_κ, depends on the maximal curvature, κ, of a certain manifold (exact definitions of z_κ and κ are given in Appendix C5.2). 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_0^2 is equal to the $(1-\alpha)$'th quantile of a central χ_1^2 random variable. As $\kappa \to \infty$, z_κ^2 approaches the $(1-\alpha)$'th quantile of a central χ_{q+1}^2 random variable. This upper limit on z_κ is used in the projection method in its classical form as popularized in econometrics by Dufour and Jasiak (2001), while the lower limit z_0 would yield size control if θ were linear in $\mathbb{E}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)']$.

When $q = 0$, the maximal curvature is zero and $\hat{C}_0^\theta$ simplifies to $\left[\hat{\theta} \pm z_0 \hat{\mathbb{V}}[\hat{\theta}]^{1/2}\right]$ where the standard error $\hat{\mathbb{V}}[\hat{\theta}]^{1/2}$ is given in the next subsection. When $q = 1$, the maximal curvature is $\hat{\kappa} = \frac{2|\lambda_1|\hat{\mathbb{V}}[\hat{b}_1]}{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}(1-\hat{\rho}^2)^{1/2}}$ 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 $\hat{\kappa}^2$ 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 useful representation of $\hat{C}_1^\theta$ is

$$\hat{C}_1^\theta = \left[\lambda_1 b_{1,-}^2 + \theta_{1,-}, \lambda_1 b_{1,+}^2 + \theta_{1,+}\right]$$

where $b_{1,\pm}$ and $\theta_{1,\pm}$ are solutions to

$$b_{1,\pm} = \hat{b}_1 \pm z_{\hat{\kappa}} \left(\hat{\mathbb{V}}[\hat{b}_1](1-\hat{a}_\pm)\right)^{1/2}$$
$$\theta_{1,\pm} = \hat{\theta}_1 - \hat{\rho}\frac{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}}(\hat{b}_1 - b_{1,\pm}) \pm z_{\hat{\kappa}} \left(\hat{\mathbb{V}}[\hat{\theta}_1](1-\hat{\rho}^2)\hat{a}_\pm\right)^{1/2}$$

for $\hat{a}_\pm = \left(1 + \left(\frac{\mathrm{sgn}(\lambda_1)\hat{\kappa} b_{1,\pm}}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}} + \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\right)^2\right)^{-1}$.

This construction is fairly intuitive. When $\hat{\rho} = 0$, the interval has endpoints that combine

$$\lambda_1 \left(\hat{b}_1 \pm z_{\hat{\kappa}} \left(\hat{\mathbb{V}}[\hat{b}_1](1-\hat{a}_\pm)\right)^{1/2}\right)^2 \quad \text{and} \quad \hat{\theta}_q \pm z_{\hat{\kappa}} \left(\hat{\mathbb{V}}[\hat{\theta}_q]\hat{a}_\pm\right)^{1/2}$$

where $\hat{a}_\pm$ estimates the fraction of $\mathbb{V}[\hat{\theta}]$ that stems from $\hat{\theta}_1$ when b_1 is one of $b_{1,\pm}$. When $\hat{\rho}$ is non-zero, $\hat{C}_1^\theta$ involves an additional rotation of $(\hat{b}_1, \hat{\theta}_1)'$. $\hat{C}_1^\theta$ can be calculated by finding the roots of a fourth order polynomial given in Appendix C5.2.

Before proposing variance estimators, we report the requirement for asymptotic validity of our inference procedure under the conditions of Theorem 1.

Lemma 4. *If* $\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}$ *and the conditions of Theorem* 1 *hold, then*

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta \in \hat{C}_q^\theta\right) \geq 1-\alpha.$$

Remark 8. When the nuisance parameters $b_1, \ldots, b_q$ are large, i.e., $\min_{\ell\in\{1,\ldots,q\}} b_\ell^2 \to \infty$, it follows from Theorem 1 and the Delta method that $\hat{C}_0^\theta = \left[\hat{\theta} \pm z_0 \hat{\mathbb{V}}[\hat{\theta}]^{1/2}\right]$ delivers size control even when q is non-zero. The interval $\hat{C}_q^\theta$ will also provide size control, but will tend to be longer (and conservative) as $z_{\hat{\kappa}} > z_0$. Note that $b_1, \ldots, b_q$ are linear combinations of β rescaled so that their estimators $\hat{b}_1, \ldots, \hat{b}_q$ have a non-vanishing variance. Thus $\min_{\ell\in\{1,\ldots,q\}} b_\ell^2 \to \infty$ will be satisfied if the corresponding unscaled linear combinations are estimated consistently and bounded away from zero. In Section 7 we illustrate that $\hat{C}_0^\theta$ can undercover in a setting where $q > 0$ and $\min_{\ell\in\{1,\ldots,q\}} b_\ell^2$ is bounded, which serves to illustrate the fragility of the Delta method.

5.2 Asymptotic Variance Estimation

We now develop an estimator of the covariance matrix that appears in Theorem 1 and is used in construction of $\hat{C}_q^\theta$. In order to explain its final form we first consider the special cases of Propositions 1 and 2.

The low rank case

For this case the relevant variance is $\mathbb{V}[\hat{b}] = \sum_{i=1}^n w_i w_i' \sigma_i^2$ and our estimator is of the Eicker-White form but uses the leave-one-out estimators $\{\hat{\sigma}_i^2\}_{i=1}^n$

$$\hat{\mathbb{V}}[\hat{b}] = \sum_{i=1}^n w_i w_i' \hat{\sigma}_i^2.$$

Unbiasedness of $\hat{\mathbb{V}}[\hat{b}]$ is immediate and consistency follows from the same set of assumptions that lead to Proposition 1.

Lemma 5. *If the conditions of Proposition* 1 *holds, then* $\mathbb{V}[\hat{b}]^{-1}\hat{\mathbb{V}}[\hat{b}] \xrightarrow{p} I_r$.

Remark 9. In the special case where $A = vv'$ for some non-random vector v, this result implies that

$$\frac{v'(\hat{\beta}-\beta)}{\sqrt{\hat{\mathbb{V}}[v'\hat{\beta}]}} \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{for} \quad \hat{\mathbb{V}}[v'\hat{\beta}] = v' S_{xx}^{-1} \left(\sum_{i=1}^{n} x_i x_i' \hat{\sigma}_i^2 \right) S_{xx}^{-1} v$$

which allows for asymptotically valid inference on linear contrasts of β in a setting with many regressors and heteroscedasticity. To derive this result we assumed that $\max_i P_{ii} \leq c$ for some $c < 1$, whereas classical versions of Eicker-White variance estimators typically require that $\max_i P_{ii} \to 0$ and Cattaneo et al. (2017) provide $\max_i P_{ii} \leq c$ for some $c \leq 1/2$ as a sufficient condition for their MINQUE-type variance estimator to yield asymptotically valid inference. Thus $\hat{\mathbb{V}}[v'\hat{\beta}]$ leads to valid inference under weaker conditions than existing versions of Eicker-White variance estimators.

The high rank case

For Proposition 2 the relevant variance is that of $\hat{\theta}$ and the U-statistic representation of $\hat{\theta}$ in (9) 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 using $\{\hat{\sigma}_i^2\}_{i=1}^n$ to form an estimator of $\mathbb{V}[\hat{\theta}]$ will in general not lead to valid inference as $\hat{\sigma}_i^2 \hat{\sigma}_\ell^2$ is not an unbiased estimator of $\sigma_i^2 \sigma_\ell^2$. Additionally, $\mathbb{V}[\hat{\theta}]$ depends on the unknown $x_\ell'\beta$ which will also have to be estimated. While $\mathbb{V}[\hat{\theta}]$ can, in principle, be estimated without bias using leave-*three*-out estimators, this approach will be computationally intractable in many settings. Moreover, in the case of Example 4, it is likely that discarding particular triples of observations will lead the mobility network to become disconnected, making it impossible to compute estimates of β.

Given these considerations, we follow a classical approach where each $\hat{\sigma}_i^2$ is smoothed using local averages. Let

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

where $\tilde{\sigma}_i^2 = \tilde{\sigma}^2(\omega_i)$, $\omega_i = (B_{ii}, P_{ii})'$, and $\tilde{\sigma}^2(\omega) = \sum_{i=1}^n \hat{\sigma}_i^2 k_i(\omega)$ for a sequence of weight functions $\{k_i\}$ such that $\sum_{i=1}^n k_i(\omega_\ell) = 1$ and $k_i(\omega_\ell) = k_\ell(\omega_i)$ for any i, ℓ. The subtraction (as opposed to addition) of the second term in the definition of $\hat{\mathbb{V}}[\hat{\theta}]$ is intentional as the use of y_ℓ in place of $x_\ell'\beta$ in the first term of $\hat{\mathbb{V}}[\hat{\theta}]$ leads to an approximate bias of $4 \sum_{i=1}^n \sum_{\ell \neq i} C_{i\ell}^2 \sigma_i^2 \sigma_\ell^2$.

Inference based on $\hat{\mathbb{V}}[\hat{\theta}]$ is asymptotically valid when: (*a*) σ_i^2 is a Lipschitz function of ω_i and

(b) the sequence of weight functions $\{k_i\}$ satisfy standard conditions which imply that $\tilde{\sigma}^2(\omega_i) - \sigma_i^2$ converge to zero (uniformly over i) in mean squared error.

Assumption 3. *(a) For some $\sigma^2(\cdot) : \mathbb{R}^2 \to \mathbb{R}_+$, $c < \infty$, and norm $\|\cdot\|$, $\sigma_i^2 = \sigma^2(\omega_i)$ and $|\sigma^2(\omega) - \sigma^2(\omega')| \leq c\|\omega - \omega'\|$ for any $\omega, \omega' \in \mathbb{R}^2$. (b) $\max_i \sum_{\ell=1}^n k_\ell(\omega_i)^2 + |k_\ell(\omega_i)|\|\omega_i - \omega_\ell\| = o(1)$.*

Lemma 6. *If the conditions of Proposition 2 and Assumption 3 hold, then $\hat{\mathbb{V}}[\hat{\theta}]/\mathbb{V}[\hat{\theta}] \xrightarrow{p} 1$.*

Remark 10. If $\tilde{\sigma}^2(\omega)$ is a locally linear kernel estimator based on a neighborhood $\mathcal{N}_\omega = \{\omega_i : \|\omega_i - \omega\| \leq h(\omega)\}$ with $|\mathcal{N}_\omega|$ nearest neighbors using a tricube kernel, then Assumption 3(b) is satisfied when the minimum number of neighbors $\min_i |\mathcal{N}_{w_i}|$ goes to infinity and the maximal neighbor distance $\max_i h(\omega_i)$ goes to zero. In our implementation, we rely on a common number of neighbors being chosen so that $\tilde{\sigma}^2(\omega)$ is rate optimal when $\sigma^2(\cdot)$ has two derivatives. We use $\|\omega_i\| = (B_{ii}^2/\sigma_B^2 + P_{ii}^2/\sigma_P^2)^{1/2}$, the standard Euclidean distance weighted by sample standard deviations, to define neighborhoods.

The general case

In Theorem 1 the relevant variance is $\Sigma_q = \mathbb{V}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)']$,

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

where $C_{i\ell q}$ is defined in Appendix C4. Our estimator of this variance reuses the ideas introduced for Propositions 1 and 2:

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

The following result shows consistency of this variance estimator.

Lemma 7. *If the conditions of Theorem 1 and Assumption 3 holds, then $\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}$.*

5.3 Choosing q

It is possible to infer q in large samples provided that Assumption 2 is adjusted slightly to include a rate condition on the eigenvalues that are small relative to their sum.

Assumption 2′. *There exist a $c > 0$, $\epsilon > 0$, and a fixed $q \in \{1, \ldots, r-1\}$ such that*

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

A "rule-of-thumb" choice of q based on Assumption 2′ is the unique $\hat{q}$ for which

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

Under Assumption 2′, $\hat{q} = q$ in sufficiently large samples provided that c_r is chosen so that $c_r r^\varepsilon \to \infty$. For instance, this condition is satisfied when c_r shrinks to zero slower than $1/\log(r)$. Note that $\frac{\lambda_{\hat{q}+1}^2}{\sum_{\ell=1}^r \lambda_\ell^2}$ can be thought of as summarizing inverse effective sample size for the weighted average $\hat{\theta}_{\hat{q}}$ of $\hat{b}_{\hat{q}+1}, \ldots, \hat{b}_r$. Our Monte Carlo study suggests good performance of the asymptotic approximations for effective sample sizes as low as 10, which is in line with statistics folklore (see, e.g., Lei, Bickel, and Karoui, 2016, page 20). We leave the study of which particular rules of thumb work best across a wide array of settings to future work.

Remark 11. For any given rule of thumb choice $\hat{q}$, one may also report the more conservative interval $\hat{C}_{\hat{q}}^\theta \cup \hat{C}_{\hat{q}+1}^\theta$. In our application we find that $\hat{C}_0^\theta$ and $\hat{C}_1^\theta$ are nearly indistinguishable in settings with $\frac{\lambda_1^2}{\sum_{\ell=1}^r \lambda_\ell^2} \leq \frac{1}{10}$, which suggest that little power may be lost from such a cautionary approach.

6 Verifying Conditions

We now revisit the examples of Section 2 and verify the conditions required to apply our theoretical results. Appendix C6 provides further details on these calculations.

Example 1. (Coefficient of determination, continued) Recall that $\theta = \sigma_{X\beta}^2 = \beta' A \beta$ where $A = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})'$ and $\tilde{A} = \frac{1}{n}\left(I_k - n S_{xx}^{-1/2} \bar{x}\bar{x}' S_{xx}^{-1/2}\right)$. Supposing Assumption 1 holds, consistency follows from Lemma 3 since $\lambda_\ell = \frac{1}{n}$ for $\ell = 1, \ldots, r$ where $r = \dim(x_i) - 1$. Thus $\text{trace}(\tilde{A}^2) = r/n^2 \leq 1/n = o(1)$. If r is fixed, then $w_i' w_i = P_{ii} - \frac{1}{n}$ and Proposition 1 applies under the "textbook" condition that $\max_i P_{ii} = o(1)$. If $r \to \infty$, then Proposition 2 applies if $\mathbb{V}[\hat{\theta}]^{-1} \max_i (\check{x}_i'\beta)^2 = o(1)$ which follows if, e.g., $\max_i \frac{1}{\sqrt{r}} \sum_{\ell=1}^n |M_{i\ell}| = o(1)$ where $M_{i\ell} = \mathbf{1}_{\{i=\ell\}} - x_i' S_{xx}^{-1} x_\ell$. Equality among all eigenvalues excludes the conditions of Theorem 1. Inspection of the proofs reveals that Assumption 1(c), $\max_i (x_i'\beta)^2 = O(1)$, can be dropped if $\max_{i,\ell} P_{ii}(x_\ell'\beta)^2 = o(1)$ when r is fixed or if $\max_{i,j} \frac{|x_j'\beta|\left(1+\sum_{\ell=1}^n |M_{i\ell}|\right)}{\sqrt{r}} = o(1)$ when $r \to \infty$.

Example 2. (Analysis of covariance, continued) With no common regressors, this is a special case of the previous example with $r = N-1$, $P_{ii} = T_{g(i)}^{-1}$ and $\check{x}_i = 0$. Assumption 1(b),(c) requires $T_g \geq 2$ and $\max_g \alpha_g^2 = O(1)$. Proposition 1 applies if N is fixed and $\min_g T_g \to \infty$, while Proposition 2 applies if $N \to \infty$. Theorem 1 cannot apply to this example.

To accomodate 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 our conclusions: Proposition 1 applies if N is fixed and $\min_g T_g \to \infty$ since $w_i'w_i \leq P_{ii} = T_{g(i)}^{-1} + O(n^{-1})$, Proposition 2 applies if $N \to \infty$ since $\sum_{\ell=1}^{n} |M_{i\ell}| = O(1)$, and Theorem 1 cannot apply since $n\lambda_\ell \in [c_1, c_2]$ for $\ell = 1, \ldots, r$ and some $c_2 \geq c_1 > 0$ not depending on n. 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 $\sigma_\alpha^2 + \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}\|x_{gt}\|^2 = O(1)$.

Example 3. (Random coefficients, continued) Consider the second moment $\theta = \frac{1}{n}\sum_{g=1}^{N} T_g \gamma_g^2$, impose Assumption 1, 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}. The N eigenvalues of $\tilde{A}$ are $\lambda_g = \frac{1}{n}\frac{1}{T_g^{-1}S_{zz,g}}$ for $g = 1, \ldots, N$ where the group indexes are ordered so that $\lambda_1 \geq \cdots \geq \lambda_N$. Consistency follows from Lemma 3 if $\lambda_1^{-1} = n\frac{S_{zz,1}}{T_1} \to \infty$. If N is fixed and $\min_g S_{zz,g} \to \infty$, then Proposition 1 applies. If $\frac{\sqrt{N}}{T_1}S_{zz,1} \to \infty$, then Proposition 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 1 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 provides the influence of γ_1 on θ and $1/\sqrt{N}$ indexes the variability of $\hat{\theta}$.

Example 4. (Two-way fixed effects, continued) In this final example, we focus on whether Proposition 2 or Theorem 1 applies. Our target parameter is the variance of firm effects $\theta = \sigma_\psi^2 = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}\left(\psi_{j(g,t)} - \bar{\psi}\right)^2$ and we restrict attention to the first-differenced setting of (7) with $J \to \infty$. The eigenvalues of $\tilde{A}$ satisfy the equality

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

where $\dot{\lambda}_1 \geq \cdots \geq \dot{\lambda}_J$ are the non-zero 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 Appendix C6 for precise definitions). E is a diagonal matrix of employer specific "churn rates", i.e., 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 3, 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 $\mathcal{L}$ by assuming that the diagonal entries of E are all equal.

The worker-firm mobility network is *strongly connected* if $\sqrt{J}\mathcal{C} \to \infty$ where $\mathcal{C} \in (0, 1]$ is the isoperimetric (or Cheeger's) constant for the mobility network (see, e.g., Mohar, 1989; Jochmans and Weidner, 2016). 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 subnetworks. 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 subnetworks. It follows from the Cheeger inequality $\dot{\lambda}_J \geq 1 - \sqrt{1 - \mathcal{C}^2}$ (Chung, 1997, Theorem 2.3) and the bound $\frac{\lambda_1^2}{\sum_{\ell=1}^J \lambda_\ell^2} \leq 4(\sqrt{J}\dot{\lambda}_J)^{-2}$ that a strongly connected network yields $q = 0$, which rules out application of Theorem 1. Furthermore, a strongly connected network is sufficient (but not necessary) for consistency of $\hat{\theta}$ as $\sum_{\ell=1}^J \lambda_\ell^2 \leq \frac{J}{n}(\sqrt{n}\dot{\lambda}_J)^{-2}$.

When $\sqrt{J}\mathcal{C}$ is bounded, the network is *weakly connected* and can contain a sufficiently thin bottleneck that a linear combination of the elements of ψ is estimated imprecisely relative to its influence on θ and the total uncertainty in $\hat{\theta}$. We illustrate this in our empirical application by considering two provinces with limited mobility between them. In this setting, the between-province difference in average firm effects is weakly identified relative to the two within-province variances of firm effects, which yields one very large eigenvalue ratio indicating that $q = 1$.

7 Application

Consider again the problem of estimating variance components in a two-way fixed effect model of wage determination. Card et al. (2018) note that plug-in wage decompositions of the sort introduced by AKM typically attribute 15%-25% of overall wage variance to variability in firm fixed effects. Given the bias and potential sampling variability associated with plug-in estimates, however, it has been difficult to infer whether firms play a significantly greater role in the determination of wage inequality in some areas than others.

In this section, we use Italian social security records to formally investigate whether the variance components that comprise the AKM decomposition differ across two provinces from the Veneto region of Northeast Italy. The first province, Rovigo, has a large share of firms in the agriculture and fishing sectors and is often viewed as a lagging area within the Veneto region (Istat, 2001). The second province, Belluno, is a wealthy area that is intensive in manufacturing and contains one of the largest clusters of eyeglass production in Europe (Whitford, 2001). The two provinces lie at opposite ends of the Veneto region (Figure A.1 provides a map) and mobility between them is rather infrequent. We examine below how this limited inter-provincial mobility influences the finite sample behavior of variance component estimates in a sample that pools the two provinces together.

7.1 Data

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 employment spell taking place in the Veneto region over the years 1975-2001 covered by the Italian social security system. For each worker, we retain the unique spell yielding the highest earnings in that year. Further detail on

our processing of these records is provided in Appendix A. We limit our sample to worker-firm spells taking place in the years 1999 and 2001, which provides us a three year horizon over which to measure job mobility.

Table 1 reports the number of person-year observations available among workers employed by firms in each province's largest connected set along with the largest connected set for the pooled sample composed of the union of the two provinces. Many firms in each sample are very weakly connected to one another: the average number of movers per firm ranges from approximately 2.0 in Rovigo to 2.5 in Belluno. Our leave-out approach requires that the firm effects remain estimable after removing any single observation. The second panel of Table 1 enforces this requirement by restricting to firms that remain connected when any single mover is dropped (see Appendix B1 for computational details).[3] Pruning the sample in this way drops roughly half of the firms but less than a third of the movers, and has little effect on the mean or variance of wages.

In the pruned "leave-out connected set" the average number of movers per firm ranges from approximately 2.8 in Rovigo to nearly 3.8 in Belluno. Our theoretical results suggest that not only the number of moves, but also their distribution throughout the mobility network, influences the behavior of variance component estimates. The leave-out connected set of the union of the two provinces is portrayed in Figure 1. As the illustration makes clear, worker mobility is much more common within than between provinces. Theorem 1 and the discussion in Section 6 show that inter-provincial bottlenecks in the mobility network can generate weak identification and non-normality, a phenomenon we explore in detail below.

7.2 Estimates

Consider the following simplified version of the AKM model introduced in Section 2:

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + \varepsilon_{gt}. \qquad (g = 1, \ldots, N,\ t = 1, 2)$$

The bottom of Table 1 reports for each sample the maximum leverage ($\max_i P_{ii}$) of any person-year observation (see Appendix B for computational details). While our pruning procedure ensures $\max_i P_{ii} < 1$, it is noteworthy that $\max_i P_{ii}$ is still quite close to one, 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

[3] In Rovigo's original leave-out connected set, both the largest eigenvalue ratio $\frac{\lambda_1^2}{\sum_{\ell=1}^{r} \lambda_\ell^2}$ and weight $\max_i \mathsf{w}_{i1}^2$ associated with the variance of firm effects were above 0.1, leading to a potential violation of the conditions in Theorem 1. To reduce $\max_i \mathsf{w}_{i1}^2$ we located the mover with the highest w_{i1}^2 and removed the stayers (i.e., non-movers) from the two firms this individual moved between. This extra pruning substantially decreased $\max_i \mathsf{w}_{i1}^2$, leading to an effective sample size for $\hat{b}_1$ of 50. Point estimates were qualitatively similar when these stayers were retained in the data.

slower than the sample size, which would lead the maximum leverage to tend to zero.

Table 2 reports three sets of estimators of the AKM variance decomposition: the naive plug-in estimator $\hat{\theta}_{\text{PI}}$ originally proposed by AKM, the homoscedasticity-corrected estimator $\hat{\theta}_{\text{HO}}$ of Andrews et al. (2008), and the leave-out estimator $\hat{\theta}$. The plug-in estimator finds that the variance of firm effects in Rovigo accounts for roughly half of the total variance of wages in that province, while in Belluno the firm effect variability accounts for only 16% of overall wage variance. Although the Belluno sample is larger, variability in firm effects account for 41% of the variance of wages in the pooled Rovigo-Belluno sample, which indicates a substantial between-province component of firm variability.

Are these patterns driven by biases attributable to estimation error? Applying the homoscedastic correction of Andrews et al. (2008) shrinks the estimated variances of firm effects by roughly 10% in Rovigo and 30% in Belluno. The leave-out estimator, in turn, yields comparably sized decreases in the estimated firm effect variance relative to the homoscedastic correction, suggesting the presence of substantial heteroscedasticity in these samples. The leave-out estimates indicate that firm effect variability accounts for 36% of the variance of wages in Rovigo but only 8% of the variance in Belluno. Because the standard errors for the estimated firm effect variances are fairly small, we can conclude with some confidence that there is much more firm effect variability present in Rovigo than Belluno.

The estimated firm effect variance in the pooled Rovigo-Belluno sample is notably less precise than the province-specific estimates, which suggests that the between-province component of variance may be weakly identified. Applying the results in Remark 9, we show in Appendix Table A.1 that the difference in person-year weighted mean firm effects between Belluno and Rovigo is 0.260 with a corresponding standard error of 0.094. Evidently, the Belluno employers pay higher wages than those in Rovigo, but there is substantial uncertainty regarding the size of this differential.

Plug-in estimates of the person effect variance suggest person effects are more dispersed in Rovigo than Belluno. Applying the homoscedastic-correction reduces the magnitude of the person effect variance in both provinces but preserves their ranking. The leave-out estimator yields additional downward corrections and the associated standard errors suggest that Rovigo does in fact have a larger person effect variance than Belluno.

Plug-in estimates of the covariance between worker and firm effects are negative in both provinces. When converted to correlations, these figures suggest there is mild negative assortative matching of workers to firms. Applying the homoscedasticity correction leads the covariances to change sign in both Rovigo and Belluno but not in the pooled sample. In all cases, however, the homoscedasticity-corrected estimates indicate very small correlations between worker and firm effects. By contrast, the leave-out estimator finds a rather strong positive correlation of 0.154 in Belluno and 0.220 in Rovigo, indicating the presence of non-trivial positive assortative matching between workers and firms. Interestingly, the leave-out estimate of worker-firm correlation in the

pooled sample is only .047, indicating that the between-province component of covariance remains negative after correction. Corroborating this interpretation, we show in Appendix Table A.1 that the difference in mean person effects between Belluno and Rovigo is -0.102 with an associated standard error of 0.094. While Belluno has higher paying firms than Rovigo, our estimates suggest Belluno may actually have lower quality workers.

Finally, we examine the overall fit of the two-way fixed effects model using the coefficient of determination. The naive plug-in R^2 estimator suggests the two-way fixed effects model explains more than 90% of wage variation in each region. Homoscedasticity-correcting the R^2 yields the adjusted R^2 measure of Theil (1961). In Rovigo, the adjusted R^2 measure indicates that the two-way fixed effects model explains roughly 92% of the variance of wages, which is quite close to the figures reported in Card et al. (2013) for the German labor market. In Belluno, however, the model is found to explain only 85% of the variance. Applying the leave-out estimator yields very minor changes in estimated explanatory power relative to the homoscedasticity-corrected estimates. In sum, the two-way fixed effects model appears to provide a very comprehensive statistical summary of wage structure in the Italian data, even after accounting for the "over-fitting" that results from estimating many parameters.

7.3 Inference

Table 3 considers more carefully the problem of conducting inference on the variance of firm effects. The top panel of Table 3 reports 95% confidence intervals that arise from assuming either $q = 0$ or $q = 1$. While the former interval employs a normal approximation, the latter allows for weak identification. We also report an estimate of the curvature parameter κ used to construct the weak identification robust interval. In both Rovigo and Belluno, κ is estimated to be quite small. Accordingly, the two sets of confidence intervals are nearly identical, suggesting a normal approximation would be accurate. In the pooled Rovigo-Belluno sample, however, we find $\kappa = 1.45$ indicating that normality is a poor approximation. Accordingly, setting $q = 1$ in this sample widens the confidence interval substantially. The fact that the two province's weak identification robust confidence intervals do not overlap implies, assuming independence, that we can reject the null hypothesis that the firm effect variances are the same in Belluno and Rovigo at the $(1-0.95^2)\times 100 = 9.75\%$ level.

Theorem 1 suggested two important diagnostics for the asymptotic behavior of our estimator are the Lindeberg statistic $\max_i \mathsf{w}_{iq}'\mathsf{w}_{iq}$ and the top eigenvalue share $\frac{\lambda_1^2}{\sum_{\ell=1}^r \lambda_\ell^2}$. The bottom panel of Table 3 reports these statistics for each sample. Rovigo has a relatively large top eigenvalue, while the pooled Rovigo-Belluno sample has an enormous top eigenvalue share of 0.88. From Theorem 1, a large top eigenvalue indicates the leave-out estimator depends heavily on the square of a particular linear combination of estimated firm effects, and will therefore exhibit a substantial χ^2 component.

The relatively small top eigenvalue found in Belluno indicates the large sample distribution of the leave-out firm effect variance estimator is, in that sample, likely to be well approximated by a normal, which is in accord with the behavior of our two empirical confidence intervals. Interestingly, the sum of squared eigenvalues is quite small in all three samples, indicating that the leave out estimator is consistent even in the pooled Rovigo-Belluno sample.

One can think of the Lindeberg statistic $\max_i \mathsf{w}_{i1}^2$ as an inverse measure of effective sample size available for estimating the linear combination of firm effects associated with the largest eigenvalue. The effective sample size in Rovigo is $\frac{1}{.02} = 50$. In Belluno, by contrast, the effective sample size is less than 4. Fortunately, the top eigenvalue share in Belluno is small, suggesting that mistakes in estimating the relevant linear combination of firm effects are not particularly consequential for inference.

7.4 Monte Carlo Experiments

We turn now to studying the finite sample behavior of our leave-out estimator of the variance of firm effects and the performance of our asymptotic inference procedures. Data were generated from the following first differenced model based on equation (7):

$$\Delta y_g = \Delta f_g' \hat{\psi} + \Delta \varepsilon_g, \qquad (g = 1, \ldots, N)$$

Here $\hat{\psi}$ gives the $J \times 1$ vector of OLS firm effect estimates, rescaled to have the province-specific leave-out variance reported in Table 2. The errors $\Delta\varepsilon_g$ were drawn independently from a student t distribution with 5 degrees of freedom. Each error was then rescaled to match the smoothed leave-out estimate $\tilde{\sigma}_i^2$ of that observation's error variance (see Appendix B for details).

For each province, we sampled from the above DGP 10,000 times holding firm assignments fixed at their realized sample values. We then applied our leave-out estimator to each simulated dataset and constructed the corresponding 95% confidence intervals. Table 4 shows that the leave-out estimator is unbiased and that the standard error estimate is also approximately unbiased. The coverage rates exhibited by the normal theory confidence interval are, in each province, close to their nominal level. By contrast, the weak identification robust confidence intervals exhibit very mild over-coverage. Evidently, Belluno's large Lindeberg statistic does not, in this case, compromise inference.

In the pooled Rovigo-Belluno sample, the normal theory confidence interval undercovers substantially, which is to be expected given the large top eigenvalue in this sample. Applying the weak identification robust interval again generates very mild over-coverage despite the fact that the effective sample size available for the top eigenvector is only $1/.0378 \approx 26$. In sum, the Monte Carlo experiments suggest confidence intervals predicated on the assumption that $q = 1$ can provide

adequate size control even when the realized mobility network exhibits very severe bottlenecks.

8 Conclusion

We proposed a new estimator of quadratic forms with applications to several areas of economics. This estimator is finite sample unbiased and consistent in the presence of heteroscedasticity and many regressors, including in circumstances where "automatic" bias correction procedures fail. A new distributional theory was developed highlighting the potential for this estimator to exhibit deviations from normality when certain linear combinations of coefficients are imprecisely estimated. We also developed a feasible inference procedure that is uniformly asymptotically valid in the presence of weakly identified nuisance parameters.

In an application to Italian worker-firm data, we demonstrated that ignoring heteroscedasticity can substantially bias conclusions about the relative contribution of workers, firms, and worker-firm sorting to wage inequality. We also found that bottlenecks in the mobility network can generate quantitatively important deviations from normality. Our inference procedure captured these deviations accurately with a choice of $q = 1$. In cases where the mobility network was strongly connected, we found that choosing $q = 0$ yielded accurate inference.

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.

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

Achlioptas, D. (2001). Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281. ACM.

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.

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

Andrews, D. W. (1988). Chi-square diagnostic tests for econometric models: theory. *Econometrica: Journal of the Econometric Society*, 1419–1453.

Andrews, I. and A. Mikusheva (2016). A geometric approach to nonlinear econometric models. *Econometrica 84*(3), 1249–1264.

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.

Angrist, J., G. Imbens, and A. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics 30*, 98–108.

Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler (2012). Estimating spillovers using panel data, with an application to the classroom. *Quantitative Economics 3*(3), 421–470.

Arellano, M. and S. Bonhomme (2011). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*(3), 987–1020.

Arriaga, R. I. and S. Vempala (1999). An algorithmic theory of learning: Robust concepts and random projection. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 616–623. IEEE.

Bonhomme, S. (2017). Econometric analysis of bipartite networks. *Mimeo*.

Bonhomme, S., T. Lamadon, and E. Manresa (2017a). Discretizing unobserved heterogeneity. *Unpublished manuscript, University of Chicago*.

Bonhomme, S., T. Lamadon, and E. Manresa (2017b). A distributional framework for matched employer employee data. *Unpublished manuscript, University of Chicago*.

Borovičková, K. and R. Shimer (2017). High wage workers work for high wage firms. Technical report, National Bureau of Economic Research.

Bryk, A. S. and S. W. Raudenbush (1992). Hierarchical linear models for social and behavioral research: Applications and data analysis methods.

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.

Card, D., F. Devicienti, and A. Maida (2014). Rent-sharing, holdup, and wages: Evidence from matched panel data. *The Review of Economic Studies 81*(1), 84–111.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics 128*(3), 967–1015.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2016). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 1–25.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2017). Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association* (just-accepted).

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.

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.

Chatterjee, S. (2008). A new method of normal approximation. *The Annals of Probability 36*(4), 1584–1610.

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.

Chung, F. R. (1997). *Spectral graph theory*. Number 92. American Mathematical Soc.

Cochran, W. G. (1980). Fisher and the analysis of variance. In *RA Fisher: An Appreciation*, pp. 17–34. Springer.

Davidson, R. and J. G. MacKinnon (1993). Estimation and inference in econometrics.

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies 82*(3), 991–1030.

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.

Drineas, P., M. Magdon, M. W. Mahoney, and D. P. Woodruff (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research 13*(Dec), 3475–3506.

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.

Efron, B. and C. Stein (1981, 05). The jackknife estimate of variance. *Ann. Statist. 9*(3), 586–596.

Finkelstein, A., M. Gentzkow, and H. Williams (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics 131*(4), 1681–1726.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*(3), 643–660.

Graham, B. S., J. Hahn, A. Poirier, and J. L. Powell (2016). A quantile correlated random coefficients panel data model. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in irregular correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152.

Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica 72*(4), 1295–1319.

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.

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.

Istat (2001). Il sistema produttivo del veneto. Technical report.

Jochmans, K. and M. Weidner (2016). Fixed-effect regressions on network data. *arXiv preprint arXiv:1608.01532*.

Karoui, N. E. and E. Purdom (2016). Can we trust the bootstrap in high-dimension? *arXiv preprint arXiv:1608.00696*.

Koutis, I., G. L. Miller, and D. Tolliver (2011). Combinatorial preconditioners and multilevel solvers for problems in computer vision and image processing. *Computer Vision and Image Understanding 115*(12), 1638–1646.

Kuh, E. (1959). The validity of cross-sectionally estimated behavior equations in time series applications. *Econometrica: Journal of the Econometric Society*, 197–214.

Lei, L., P. J. Bickel, and N. E. Karoui (2016). Asymptotics for high dimensional regression m-estimates: Fixed design results. *arXiv preprint arXiv:1612.06358*.

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.

Mohar, B. (1989). Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B 47*(3), 274–291.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics 32*(3), 385–397.

Newey, W. K. and J. R. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Phillips, G. D. A. and C. Hale (1977). The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review*, 219–228.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 1403–1430.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological) 11*(1), 68–84.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association 65*(329), 161–172.

Raudenbush, S. and A. S. Bryk (1986). A hierarchical model for studying school effects. *Sociology of education*, 1–17.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics 116*(2), 681–704.

Scheffe, H. (1959). The analysis of variance. Technical report.

Searle, S. R., G. Casella, and C. E. McCulloch (2009). *Variance components*, Volume 391. John Wiley & Sons.

Serafinelli, M. (2017). Good firms, worker flows and local productivity.

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.

Silver, D. W. (2016). Essays on labor economics and health care.

Sølvsten, M. (2017). Robust estimation with many instruments. *Unpublished manuscript, University of Wisconsin - Madison*.

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. Von Wachter (2017). Firming up inequality. Technical report, National Bureau of Economic Research.

Sorkin, I. (2017). Ranking firms using revealed preference. Technical report, National Bureau of Economic Research.

Spielman, D. A. and N. Srivastava (2011). Graph sparsification by effective resistances. *SIAM Journal on Computing 40*(6), 1913–1926.

Swamy, P. A. (1970). Efficient inference in a random coefficient regression model. *Econometrica: Journal of the Econometric Society*, 311–323.

Verdier, V. (2016). Estimation and inference for linear models with two-way unobserved heterogeneity and sparsely matched data. Technical report, Mimeo.

Whitford, J. (2001). The decline of a model? challenge and response in the italian industrial districts. *Economy and society 30*(1), 38–65.

Woodbury, M. A. (1949). The stability of out-input matrices. *Chicago, IL 9*.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research 20*(7), 557–585.

Table 1: Comparing Samples and Places

	Rovigo	**Belluno**	**Rovigo - Belluno**
	[1]	[2]	[3]
Largest Connected Set			
Number of Observations	43,330	63,462	106,964
Number of Movers	5,061	7,921	13,022
Number of Firms	2,579	3,131	5,732
Mean Log Daily Wage	4.6089	4.7482	4.6917
Variance Log Daily Wage	0.1560	0.1256	0.1427
Leave Out Sample (Pruned)			
Number of Observations	32,848	56,044	89,666
Number of Movers	3,531	6,414	9,972
Number of Firms	1,282	1,684	2,974
Mean Log Daily Wage	4.6015	4.7636	4.7047
Variance Log Daily Wage	0.1674	0.1245	0.1465
Maximum Leverage (P_{ii})	0.9241	0.9085	0.9236

Note: Data in each column corresponds to person year observations in the years 1999 and 2001 belonging to a given province in Veneto, where the last column represents the union of the Rovigo and Belluno provinces. Largest connected set represents the largest sample in which all the associated firms are connected. The leave out sample is the largest connected set such that every firm remains connected after removing any given edge (mover), see Appendix B for details. We further pruned this sample by removing any stayer belonging to the firms associated with the mover with the highest lindeberg condition. A mover is defined as a worker who switched firm between the year 1999 and 2001. Statistics on log daily wages are person-year weighted. Source: VWH dataset.

Table 2: Variance Decomposition

	Rovigo	**Belluno**	**Rovigo - Belluno**
	[1]	[2]	[3]
Variance of Log Wages	0.1674	0.1245	0.1465
Variance of Firm Effects			
Plug in (AKM)	0.0831	0.0198	0.0607
Homoscedatic Correction	0.0722	0.0136	0.0538
Leave Out	0.0609	0.0103	0.0442
	(0.0083)	(0.0011)	(0.0110)
Covariance Firm, Worker Effects			
Plug in (AKM)	-0.0072	-0.0039	-0.0126
Homoscedatic Correction	0.0030	0.0018	-0.0038
Leave Out	0.0138	0.0046	0.0028
	(0.0043)	(0.0009)	(0.0076)
Variance of Worker Effects			
Plug in (AKM)	0.0926	0.1035	0.1032
Homoscedatic Correction	0.0758	0.0883	0.0859
Leave Out	0.0647	0.0853	0.0792
	(0.0043)	(0.0011)	(0.0038)
Correlation of Worker, Firm Effects			
Plug in (AKM)	-0.0821	-0.0863	-0.1593
Homoscedatic Correction	0.0409	0.0511	-0.0555
Leave Out	0.2202	0.1538	0.0469
Coefficient of Determination			
Plug in (AKM)	0.9637	0.9280	0.9463
Homoscedatic Correction	0.9213	0.8490	0.8850
Leave Out	0.9153	0.8414	0.8797

Note: Results for each province computed using the leave out connected sample described in the bottom panel of Table 1. Numbers in parentheses refer to asymptotic standard errors calculated using the ``high rank'' variance estimator described in Section 5.2. All variance components are person-year weighted.

Table 3: Inference on Variance of Firm Effects

	Rovigo	**Belluno**	**Rovigo - Belluno**
	[1]	[2]	[3]
Variance of Firm Effects			
Leave out estimate	0.0609	0.0103	0.0442
	(0.0083)	(0.0011)	(0.0110)
95% Confidence Intervals - Strong id (q=0)	[0.0446; 0.0771]	[0.0081; 0.0125]	[0.0226; 0.0658]
95% Confidence Intervals - Weak id (q=1)	[0.0455; 0.0795]	[0.0081; 0.0127]	[0.0288; 0.0786]
Curvature ($\hat{\kappa}$)	0.1792	0.1372	1.4448
Diagnostics			
Eigenvalue Ratio - 1	0.1062	0.0465	0.8866
Eigenvalue Ratio - 2	0.0623	0.0439	0.0132
Eigenvalue Ratio - 3	0.0499	0.0348	0.0081
Lindeberg Condition ($\max_i w_{i1}^2$)	0.0200	0.2681	0.0378
Sum of Squared Eigenvalues	0.0006	0.0002	0.0001

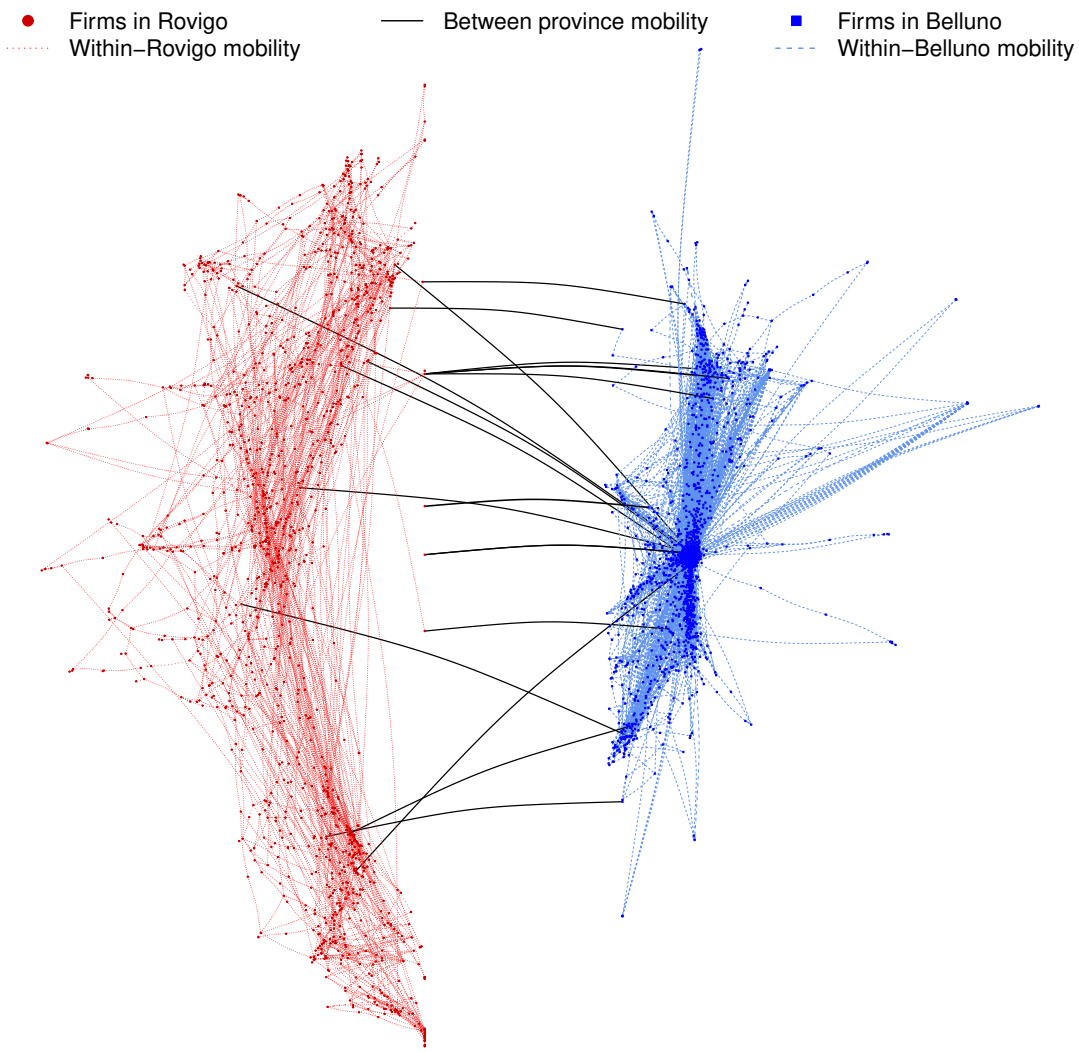
Note: Results for each province computed using the leave out connected sample described in bottom panel of Table 1. To compute the standard error, we fit a local linear estimator using a tricube kernel with nearest neighbors with bandwidth set to n^(-1/3) where n is the number of person-year observations, see Appendix B for details. Lindeberg condition and weak id confidence intervals calculated assuming that q=1. Curvature reports the maximal curvature defined in Section 6.1, see Appendix C5.2 for further details. Critical values to compute the weak identification confidence intervals of Andrews and Mikusheva (2016) based on 10,000 simulations. Eigenvalue ratio - 1 is equal to the ratio of the squared largest eigenvalue relative to the sum of all squared eigenvalues. Similarly for Eigenvalue Ratio - 2 and Eigenvalue Ratio - 3 which use the second and third largest eigenvalues respectively.

Table 4: Montecarlo Results for the Variance of Firm Effects

	[1] **Rovigo**	[2] **Belluno**	[3] **Rovigo - Belluno**
True Variance of the Firm Effects	0.0609	0.0103	0.0442
Mean, Standard deviation across Simulations			
Leave Out	0.0608	0.0103	0.0443
	(0.0073)	(0.0010)	(0.0116)
Plug-in (AKM)	0.0841	0.0196	0.0619
	(0.0073)	(0.0010)	(0.0116)
Homoscedatic Correction	0.0735	0.0134	0.0524
	(0.0073)	(0.0010)	(0.0116)
Mean estimated Standard Error	0.0074	0.0010	0.0108
Coverage Rate at 95%			
Leave Out - Strong Id (q=0)	0.9479	0.9469	0.8535
Leave Out - Weak Id (q=1)	0.9634	0.9701	0.9736

Note: Monte Carlo results based on the observed network structure of the listed provinces in the years 1999 and 2001. Data were generated by summing the plug-in firm effects estimates (rescaled to match the leave-out variance estimate) and a t-distributed error with 5 degrees of freedom and observation specific variance equal to the smoothed estimate described in Appendix B, see Section 7 for details. Leave Out, AKM, Homoscedastic correction report the average of the estimate of the firm effects across simulations for the three different methodologies. ``Leave Out - Strong Id" builds a confidence interval using the leave out estimate of the variance of firms effects, the normal distribution quantile and the estimated standard error of the variance of firm effects under the high rank case described in Section 5.2 ``Leave Out - Weak Id" builds a confidence interval using the Andrews and Mikusheva (2016) methodology with q=1. Number of Monte Carlo draws is fixed at 10,000.

Figure 1: Realized Mobility Network: Rovigo and Belluno



Note: This figure provides a visualization of the design matrix S_{xx} for the pruned Rovigo-Belluno sample considered in the application (see Table 1 for reference). The graph is plotted in the statistical software R using the *igraph* package and the large-scale graph layout (DrL).

Appendix A: Data and Additional Results

Here we describe the data used for our application and report some additional results.

A1 Veneto Workers History

Our data come from the Veneto Workers History (VWH) file, which provides social security based earnings records on annual job spells for all workers employed in the Italian region of Veneto at any point between the years 1975 and 2001. Each job-year spell in the VWH lists a start date, an end date, the number of days worked that year, and the total wage compensation received by the employee in that year. The earnings records are not top-coded. We also observe the gender of each worker and several geographic variables indicating the location of each employer. See Card, Devicienti, and Maida (2014) and Serafinelli (2017) for additional discussion and analysis of the VWH.

To construct the person-year panel used in our analysis, we follow closely the sample selection procedures described in Card, Heining, and Kline (2013). First, we drop employment spells in which the worker's age lies outside the range 20-60. The average worker in this sample has 1.21 jobs per year. To generate unique worker-firm assignments in each year, we restrict attention to spells associated with "dominant jobs" where the worker earned the most in each corresponding year. From this person-year file, we then exclude workers that (i) report a daily wage less than 5 real euros or have zero days worked (1.5% of remaining person-year observations) (ii) report a log daily wage change one year to the next that is greater than 1 in absolute value (6%) (iii) are employed in the public sector (10%) or (iv) have more than 10 jobs in any year or that have gender missing (0.1%).

Appendix Table A.1: Provincial Differences in Mean Effects

Firm Effects	
Avg. Firm Effects (Belluno)	-0.0189
Avg. Firm Effects (Rovigo)	-0.2787
Difference	0.2598
	(0.0941)
Lindeberg Condition ($\max_i w_{i1}^2$)	0.0381
Person Effects	
Avg. Person Effects (Belluno)	4.7823
Avg. Person Effects (Rovigo)	4.8854
Difference	-0.1020
	(0.0941)
Lindeberg Condition ($\max_i w_{i1}^2$)	0.0381

Note: This table compares average firm and person effects across provinces in the pooled Rovigo-Belluno sample. Standard error for the difference between the two means is reported in parentheses and computed as described in Remark 9. Lindeberg condition computed assuming A=vv' where v is such that v'*β returns the person-year weighted difference in fixed effect means across the two provinces. See text for details. Source: VWH dataset.

Figure A1: The seven provinces of Veneto

Appendix B: Computation

In this Appendix we describe computation of the leave out estimator $\hat{\theta}$, with an emphasis on the application to two-way fixed effects models discussed in Section 7.

B1 Leave One Out Connected Set

Computing $\hat{\theta}$ requires $P_{ii} < 1$ (see Assumption 1). In the two-way fixed effects model of Section (7.2), this condition requires that the bipartite graph formed by worker-firms links remains connected when any one worker is removed. This condition fails if a firm has only one worker that either left or joined the firm across the two periods.

Below we describe an algorithm that prunes the data to ensure that $P_{ii} < 1$. The input data is a connected bipartite graph $\mathcal{G}$ where vertices are represented by workers and firms and edges correspond to the realization of a match between a worker and a firm (see Jochmans and Weidner, 2016; Bonhomme, 2017, for discussion). In practice, one typically starts with a $\mathcal{G}$ corresponding to the largest connected component of the graph (see, e.g., Card et al., 2013).

Algorithm 1 Leave One Out Connected Set

function PruningNetwork(G) ▷ $\mathcal{G} \equiv$ Connected graph from bipartite network of firms and workers
  $a = 1$
  while $a > 0$ **do**
    $\mathcal{G}^{bad} = \emptyset$.
    for $g = 1, \dots N$ **do**
      Add g to $\mathcal{G}^{bad}$ if removal of worker g from $\mathcal{G}$ disconnects the resulting graph.
    end for
    $a = |\mathcal{G}^{bad}|$.
    Update $\mathcal{G}$ by finding the largest connected set after removing workers in $\mathcal{G}^{bad}$.
  end while
end function

The output of this algorithm is a bipartite graph where removal of any given worker does not break the connectedness of the graph. To find such graph, the algorithm iteratively searches for, and then removes, workers that represent articulation points in $\mathcal{G}$. This can be done very efficiently using the Boost Graph Library available for Matlab.

B2 Leave One Out Matrices

Leave one out estimation hinges on computation of the leverage scores $P_{ii} = x_i' S_{xx}^{-1} x_i \ \forall i = 1, \dots n$. Fast computation of these scores is an active area of research in computer science (see, e.g., the discussion in Drineas et al., 2012). We use recent advances in this area to illustrate how these scores can be computed efficiently in two-way fixed effects models.

Without loss of generality, we can write the model of Section 7.2 as

$$y_i = x_i'\beta + \varepsilon_i$$

where $x_i = (d_i', -\dot{f}_i')'$, $\dot{f}_i = (\mathbf{1}_{j(g,t)=0}, \dots, \mathbf{1}_{j(g,t)=J})'$, $\beta = (\alpha', -\psi')'$ and $\psi = (\psi_0, \dots, \psi_J)$.

It is easy to verify that in this case $S_{xx} = \dot{\mathcal{L}}$ where $\dot{\mathcal{L}}$ is the weighted Laplacian associated with the bipartite graph $\mathcal{G}$ formed by workers and firms. This implies that:

$$\begin{aligned} P_{ii} &= x_i' S_{xx}^{\dagger} x_i \\ &= \dot{\mathcal{L}}^{\dagger}_{g,g} + \dot{\mathcal{L}}^{\dagger}_{N+j(g,t),N+j(g,t)} - 2\dot{\mathcal{L}}^{\dagger}_{g,N+j(g,t)} \\ &= (e_g - e_{N+j(g,t)})' \dot{\mathcal{L}}^{\dagger} (e_g - e_{N+j(g,t)}) \\ &= (e_g - e_{N+j(g,t)})' \dot{\mathcal{L}}^{\dagger} \dot{\mathcal{L}} \dot{\mathcal{L}}^{\dagger} (e_g - e_{N+j(g,t)}) \\ &= ||X \dot{\mathcal{L}}^{\dagger} (e_g - e_{N+j(g,t)})||^2 \end{aligned}$$

where e_i represents the elementary unit vector with a coordinate of 1 in position i, X stacks all x_i''s and $S_{xx}^{\dagger}$ is the Moore-Penrose inverse of the Laplacian matrix. The last line of the above expression reveals that P_{ii} represents a particular pairwise distance between column vectors of the matrix $Z = X\dot{\mathcal{L}}^{\dagger}$. Obtaining these distances however involves computation of a very large block diagonal linear system that has a total of $k \times n$ unknowns, where here $k = N + J$, i.e. the total number of workers and firms observed in the data.[4]

There are (at least) two possible approaches to solving this system. The first is to parallelize computation of P_{ii} (and B_{ii}) across different cores. We pursue this idea when computing estimates for the three provinces shown in our application. The second idea, which is more suitable when working with millions of workers and firms, is to work with a randomized "sketch" of the matrix Z and use this sketch to estimate the differences between rows of the matrix Z. This is the intuition behind the Johnson-Lindenstrauss Lemma (JLL) presented below

Lemma B2.1. *(Achlioptas, 2001). Given fixed vectors $z_1, \dots, z_k \in \mathbb{R}^n$ and $\epsilon > 0$, let $\mathcal{Q} \in \mathbb{R}^{p \times n}$ be a random Rademacher matrix with entries $\pm 1/\sqrt{p}$ with $p \geq 24\log(k)/\epsilon^2$. Then with probability at*

[4] An alternative would be to consider the QR decomposition of the matrix S_{xx}. However, as described in Drineas et al. (2012), the QR decomposition runs in $O(nk^2)$ time and may therefore become intractable in large datasets.

least $1 - 1/k$

$$(1-\epsilon)||z_\kappa - z_{\kappa'}||^2 \leq ||\mathcal{Q}z_\kappa - \mathcal{Q}z_{\kappa'}||^2 \leq (1+\epsilon)||z_\kappa - z_{\kappa'}||^2$$

for all pairs (κ, κ') *with* $\kappa, \kappa' \in \{1, \ldots, k\}$.

The JLL implies that we can ϵ-approximate all the statistical leverages in our bipartite graph by solving only a logarithmic number (p) of linear systems. Algorithm 2 below is taken from Spielman and Srivastava (2011) and illustrates how to approximate the statistical leverages associated with the model of Section (7.2). To implement the solution step listed in row 7, we take advantage of the "CMG" solver of Koutis et al. (2011) for symmetric diagonally dominant linear systems. Computing $\approx 234,000$ firm effects and $\approx 2,200,000$ worker effects took approximately 11 seconds with the CMG solver on a 64 core machine with 256 GB of dedicated RAM. By contrast, using the method suggested in Card et al. (2013) took approximately 34 seconds.

Algorithm 2 Fast Approximation of Statistical Leverages

1: **function** LEVERAGE(x,ϵ)
2: Let $p = \frac{24 \log k}{\epsilon^2}$.
3: Construct $\mathcal{Q}$ as a random $\pm 1/\sqrt{p}$ Rademacher matrix of dimensions $p \times n$.
4: Compute $\Upsilon = \mathcal{Q}X$.
5: Let ξ_κ denote the κ'th row of Υ.
6: **for** $\kappa = 1, \ldots p$ **do**
7: Solve the system: $\dot{\mathcal{L}}\tilde{z}_\kappa = \xi'_\kappa$
8: **end for**
9: Build $\tilde{Z} = (\tilde{z}'_1, \ldots, \tilde{z}'_p)$
10: Approximate each P_{ii} as: $||\tilde{Z}(e_g - e_{N+j(g,t)})||^2$
11: **end function**

Our procedure also requires computation of $B_{ii} = x_i' S_{xx}^{-1} A S_{xx}^{-1} x_i$. When A is used to estimate variance components, then we can rewrite B_{ii} as

$$B_{ii} = (e_g - e_{N+j(g,t)})' \dot{\mathcal{L}}^\dagger A \dot{\mathcal{L}}^\dagger (e_g - e_{N+j(g,t)}) = ||A^{1/2} \dot{\mathcal{L}}^\dagger (e_g - e_{N+j(g,t)})||^2$$

Hence, one can approximate B_{ii} via a simple modification of Algorithm 2. When A is used to estimate covariance components, we can rewrite A as in Lemma 3, that is $A = A_1' A_2$. Note that a simple corollary of the JLL is that inner products are preserved under random projections (see for instance Corollary 2 in Arriaga and Vempala, 1999).

We conclude by noting that our discussion has focused on the special case where x_i includes only firm and worker indicators. When controls are included in the model, it is possible to obtain the leverage scores in an efficient way by applying the algorithm of Drineas et al. (2012), which generalizes Algorithm 2 to a setting with arbitrary design matrix S_{xx}.

B2.1 Quality of The Approximation

Table B1 reports estimates of the variance of firm effects in three samples of different sizes belonging to the VWH dataset. Two computational methods are considered: one based on Algorithm 2 and one that parallelizes computation of (P_{ii}, B_{ii}) across multiple cores and provides an exact solution.

Overall, both the maximum leverage and the leave one out estimate of the variance of firm effects based on random projections turns out to be very close to their exact counterparts. Importantly, both the quality of the approximation to the variance of the firm effects and the computation time saved relative to the exact method appear to improve as we estimate the model in larger bipartite networks.

B2.2 Computation of Standard Errors

Here we describe the standard errors reported in Table 3. To compute $\tilde{\sigma}_i^2$ and therefore $\hat{\mathbb{V}}[\hat{\theta}]$, we fit a local linear regression of the leave one out variances $\hat{\sigma}_i^2$ on normalized values of (B_{ii}, P_{ii}). We used a tricube kernel and a common bandwidth of $n^{-1/3}$ neighbors. This is performed in Matlab using the `lfit` routine. In practice, we find that different choices of the kernel and/or bandwidth deliver very similar results.

Table B1: Evaluating Computation Methods

	Belluno-Rovigo [2]	**Venice** [2]	**Veneto** [3]
Leave One Out Sample			
Number of Observations	351,029	736,362	4,512,718
Number of Movers	26,372	49,200	370,287
Number of Firms	6,218	12,447	76,971
Time to compute P_{ii} and B_{ii} (seconds)			
Exact Method	88	517	34,969
Algorithm 2	30	75	679
Variance of firm effects			
Exact Method	0.0300	0.0361	0.0293
Algorithm 2	0.0298	0.0361	0.0293
Maximum Leverage			
Exact Method	0.7028	0.7385	0.7807
Algorithm 2	0.7030	0.7529	0.7958

Note: Each column represents data taken from a different set of provinces in Veneto. Column 1 is the union of the provinces in Belluno-Rovigo in the years 1997-2001 (T=5). Column 2 is the province of Venice in the years 1997-2001. Column 3 is the entire region of Veneto observed in the years 1997-2001. Exact method refers to an algorithm that parallelizes computation of P_{ii} and B_{ii} across multiple cores. Algorithm 2 uses the Johnson Lindestrauss Lemma to find these two terms via a simulation method setting ε=0.01, see Appendix B for details. Calculations computed on 64 cores machine with 256 GB of dedicated memory. Source: VWH dataset.

Appendix C: Proofs

The following contains all technical details and proofs that where left out of the main paper. All material is presented in the order it appears in the main paper and under the same headings.

C1 Unbiased Estimation of Variance Components

Lemma C1.1. *It follows from the Sherman-Morrison-Woodbury formula that the two representations of $\hat{\theta}$ given in* (1) *and* (2) *are numerically identical, i.e., that*

$$\hat{\beta}' A \hat{\beta} - \sum_{i=1}^{n} B_{ii} \hat{\sigma}_i^2 = \sum_{i=1}^{n} y_i \tilde{x}_i' \hat{\beta}_{-i}$$

whenever S_{xx} has full rank and $\max_i P_{ii} < 1$.

Proof. The Sherman-Morrison-Woodbury formula states that if S_{xx} has full rank and $P_{ii} < 1$, then

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

Furthermore, we have that $\tilde{x}_i' S_{xx}^{-1} x_i = x_i S_{xx}^{-1} A S_{xx}^{-1} x_i = B_{ii}$ so

$$\begin{aligned} y_i \tilde{x}_i' \hat{\beta}_{-i} &= y_i \tilde{x}_i' \left(S_{xx} - x_i x_i'\right)^{-1} \sum_{\ell \neq i} x_\ell y_\ell = y_i \tilde{x}_i' S_{xx}^{-1} \sum_{\ell \neq i} x_\ell y_\ell + \frac{y_i \tilde{x}_i' S_{xx}^{-1} x_i x_i' S_{xx}^{-1}}{1 - x_i' S_{xx}^{-1} x_i} \sum_{\ell \neq i} x_\ell y_\ell \\ &= y_i \tilde{x}_i' \hat{\beta} - B_{ii} y_i^2 + y_i B_{ii} \underbrace{x_i' \frac{S_{xx}^{-1}}{1 - x_i' S_{xx}^{-1} x_i} \sum_{\ell \neq i} x_\ell y_\ell}_{= x_i' \hat{\beta}_{-i}} = y_i \tilde{x}_i' \hat{\beta} - B_{ii} y_i (y_i - x_i' \hat{\beta}_{-i}) \end{aligned}$$

where the last expression equals $y_i \tilde{x}_i' \hat{\beta} - B_{ii} \hat{\sigma}_i^2$. This finishes the proof since $\hat{\beta}' A \hat{\beta} = \sum_{i=1}^n y_i \tilde{x}_i' \hat{\beta}$. In the above the Sherman-Morrison-Woodbury formula was also used to establish that

$$x_i' \hat{\beta}_{-i} = x_i' \left(S_{xx} - x_i x_i'\right)^{-1} \sum_{\ell \neq i} x_\ell y_\ell = x_i' \frac{S_{xx}^{-1}}{1 - x_i' S_{xx}^{-1} x_i} \sum_{\ell \neq i} x_\ell y_\ell,$$

and from this it follows that $y_i - x_i' \hat{\beta}_{-i} = \dfrac{y_i - x_i' \hat{\beta}}{1 - P_{ii}}$ as claimed in the paper. □

C1.1 Relation To Existing Approaches

Next we show that the bias of $\hat{\theta}_{\text{HO}}$ is a function of the covariation between σ_i^2 and (B_{ii}, P_{ii}).

Lemma C1.2. *The bias of $\hat{\theta}_{HO}$ is*

$$\sigma_{nB_{ii},\sigma_i^2} + S_B \frac{n}{n-k} \sigma_{P_{ii},\sigma_i^2}$$

where

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2, \quad S_B = \sum_{i=1}^n B_{ii}, \quad \sigma_{nB_{ii},\sigma_i^2} = \sum_{i=1}^n B_{ii}(\sigma_i^2 - \bar{\sigma}^2), \quad \sigma_{P_{ii},\sigma_i^2} = \frac{1}{n}\sum_{i=1}^n P_{ii}(\sigma_i^2 - \bar{\sigma}^2).$$

Proof. Since $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^n (y_i - x_i'\hat{\beta})^2 = \frac{1}{n-k}\sum_{i=1}^n \sum_{\ell=1}^n M_{i\ell}\varepsilon_i\varepsilon_\ell$ we get that

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_{\mathrm{HO}}] - \theta &= \sum_{i=1}^n B_{ii}\sigma_i^2 - \left(\sum_{i=1}^n B_{ii}\right)\frac{1}{n-k}\sum_{i=1}^n M_{ii}\sigma_i^2 \\
&= \sum_{i=1}^n B_{ii}(\sigma_i^2 - \bar{\sigma}^2) - S_B\frac{1}{n-k}\sum_{i=1}^n M_{ii}(\sigma_i^2 - \bar{\sigma}^2) \\
&= \sigma_{nB_{ii},\sigma_i^2} + S_B\frac{n}{n-k}\sigma_{P_{ii},\sigma_i^2}.
\end{aligned}$$

□

C1.2 Finite Sample Properties

Here we provide a restatement and proof of Lemmas 1 and 2 together with a characterization of the finite sample distribution of $\hat{\theta}$ which was excluded from the main text.

Lemma C1.3. *Recall that $\theta^* = \hat{\beta}'A\hat{\beta} - \sum_{i=1}^n B_{ii}\sigma_i^2$.*

1. *If $\max_i P_{ii} < 1$, then $\mathbb{E}[\hat{\theta}] = \theta$.*
2. *If $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, then $\theta^* = \sum_{\ell=1}^r \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell]\right)$ and $\hat{b} \sim \mathcal{N}\left(b, \mathbb{V}[\hat{b}]\right)$ where*

$$b = Q'S_{xx}^{1/2}\beta \quad \text{and} \quad \mathbb{V}[\hat{b}] = \sum_{i=1}^n w_i w_i' \sigma_i^2.$$

3. *If $\max_i P_{ii} < 1$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, then $\hat{\theta} = \sum_{\ell=1}^{r_C} \lambda_\ell(C)\left(z_\ell^2 - V_{\ell\ell}\right)$ where $\mathcal{Z} \sim \mathcal{N}(\mu, V)$, $\mu = Q_C'X\beta$, $V = Q_C'\Omega Q_C$, $C = (C_{i\ell})_{i,\ell}$, $\Omega = diag(\sigma_1^2, \ldots, \sigma_n^2)$, and $C = Q_C D_C Q_C'$ is a spectral decomposition of C such that $D_C = diag(\lambda_1(C), \ldots, \lambda_{r_C}(C)$ and r_C is the rank of C.*

Proof of Lemma C1.3. First note that $\hat{\beta}'A\hat{\beta} = \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell}y_i y_\ell$ and $\hat{\sigma}_i^2 = y_i(y_i - x_i'\hat{\beta}_{-i}) =$

$y_i M_{ii}^{-1} \sum_{\ell=1}^n M_{i\ell} y_\ell$, so

$$\begin{aligned}\hat{\theta} &= \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} y_i y_\ell - B_{ii} M_{ii}^{-1} M_{i\ell} y_i y_\ell \\ &= \sum_{i=1}^n \sum_{\ell=1}^n \left(B_{i\ell} - 2^{-1} M_{i\ell} \left(B_{ii} M_{ii}^{-1} + B_{\ell\ell} M_{\ell\ell}^{-1} \right) \right) y_i y_\ell = \sum_{i=1}^n \sum_{\ell \neq i} C_{i\ell} y_i y_\ell.\end{aligned}$$

The errors are mean zero and uncorrelated across observations, so

$$\mathbb{E}[\hat{\theta}] = \sum_{i=1}^n \sum_{\ell \neq i} C_{i\ell} x_i' \beta x_\ell' \beta = \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x_i' \beta x_\ell' \beta - B_{ii} M_{ii}^{-1} M_{i\ell} x_i' \beta x_\ell' \beta = \theta,$$

since $\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x_i x_\ell' = A$ and $\sum_{\ell=1}^n M_{i\ell} x_\ell = 0$. This shows the first claim of the lemma.

Recall the spectral decomposition $\tilde{A} = QDQ'$ and definition that $\hat{b} = Q' S_{xx}^{1/2} \hat{\beta}$ which satisfies that $\hat{b} \sim \mathcal{N}(\mathbb{E}[\hat{b}], \mathbb{V}[\hat{b}])$ when $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. We have that

$$\theta^* = \sum_{\ell=1}^r \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell] \right)$$

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

$$\sum_{i=1}^n B_{ii} \sigma_i^2 = \operatorname{trace}(B\Omega) = \operatorname{trace}(A \mathbb{V}[\hat{\beta}]) = \operatorname{trace}(D \mathbb{V}[\hat{b}]) = \sum_{\ell=1}^r \lambda_\ell \mathbb{V}[\hat{b}_\ell].$$

where $B = (B_{i\ell})_{i,\ell}$. This shows the second claim of the lemma.

The matrix C is is well-defined as $\min_i M_{ii} > 0$. Define $\hat{y} = Q_C'(y_1, \ldots, y_n)'$ which satisfies that $\hat{y} \sim \mathcal{N}(\mu, V)$ when $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. As for the second claim we have that

$$\hat{\theta} = y' C y = \hat{y}' D_C \hat{y} = \sum_{\ell=1}^{r_C} \lambda_\ell(C) \hat{y}_\ell^2,$$

so the third claim follows from the observation that $C_{ii} = 0$ for all i, so that $\sum_\ell \lambda_\ell(C) V_{\ell\ell} = \operatorname{trace}(C\Omega) = 0$. □

C1.3 Consistency

The next result provides a restatement and proof of Lemma 3.

Lemma C1.4. *1. If A is positive semi-definite, (i) $\theta = O(1)$,*

$$(ii)\ \textit{trace}(\tilde{A}^2) = \sum_{\ell=1}^{r} \lambda_\ell^2 = o(1),$$

and Assumption 1 holds, then $\hat{\theta} - \theta \xrightarrow{p} 0$.

2. If A is non-definite then write $A = A_1' A_2$ for some A_1, A_2. If $\Theta_\ell = \beta' A_\ell' A_\ell \beta$ satisfies (i) and (ii) for $\ell = 1, 2$, then $\hat{\theta} - \theta \xrightarrow{p} 0$.

Proof of Lemma 3. Suppose that A is positive semi-definite. The difference between $\hat{\theta}$ and θ is

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

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

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

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

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

Finally, the variance of the third term is

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

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

When A is non-definite, we write $A = A_1' A_2$ and note that

$$\beta' A S_{xx}^{-1} A \beta = \beta' A_1' A_2 S_{xx}^{-1} A_2' A_1 \beta \leq \Theta_1 \lambda_1(\tilde{A}_2) \quad \text{and} \quad \text{trace}(\tilde{A}^2) \leq \text{trace}(\tilde{A}_1^2)^{1/2} \text{trace}(\tilde{A}_2^2)^{1/2}$$

where $\tilde{A}_\ell = S_{xx}^{-1/2} A_k' A_k S_{xx}^{-1/2}$ for $\ell = 1, 2$ and $\lambda_1(\tilde{A}_2)$ is the largest eigenvalue of $\tilde{A}_2$. Thus consistency of $\hat{\theta}$ follows from $\Theta_1 = O(1)$, $\text{trace}(\tilde{A}_1^2) = o(1)$, and $\text{trace}(\tilde{A}_2^2) = O(1)$. □

C2 Examples

All mathematical discussions of the examples are collected in C6.

C3 Comparison to Jackknife Estimators

For this special case of example 2 we have that $A = \frac{I_N}{N}$ and $S_{xx} = TI_N$ so that $\tilde{A} = \frac{I_N}{NT}$ and $\text{trace}(\tilde{A}^2) = \frac{1}{NT^2} = o(1)$ which implies consistency of $\hat{\theta}$. Similarly we have that the bias of $\tilde{\theta}$ is

$$\frac{1}{n}\sum_{g=1}^{N} T_g \mathbb{V}[\hat{\alpha}_g] = \frac{1}{n}\sum_{g=1}^{N} \sigma^2 = \frac{\sigma^2}{T} \qquad \text{where } \hat{\alpha}_g = \frac{1}{T_g}\sum_{t=1}^{T_g} y_{gt}.$$

The same types of calculations lead to the other biases reported in the paper.

For this special case of example 3 we have that $A = \begin{bmatrix} 0 & 0 \\ 0 & \frac{I_N}{N} \end{bmatrix}$ and $S_{xx} = \begin{bmatrix} TI_N & 0 \\ 0 & I_N \sum_{t=1}^{T} x_t^2 \end{bmatrix}$ which implies that $\text{trace}(\tilde{A}^2) = \frac{1}{N\left(\sum_{t=1}^{T} x_t^2\right)^2} = o(1)$ and therefore consistency of $\hat{\theta}$. Similarly we have that the bias of $\tilde{\theta}$ is

$$\frac{1}{n}\sum_{g=1}^{N} T_g \mathbb{V}[\hat{\delta}_g] = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2} \qquad \text{where } \hat{\delta}_g = \frac{\sum_{t=1}^{T_g} x_t y_{gt}}{\sum_{t=1}^{T} x_t^2}.$$

The same types of calculations lead to the other biases reported in the paper. Now for the numerical example

$$(x_1, x_2, \ldots, x_T) = (-1, 2, 0, \ldots, 0, -1)$$

we have that

$$\sum_{t=1}^{T} x_t^2 = 6, \qquad \sum_{s \neq t} (x_s - \bar{x}_{-t})^2 = \begin{cases} 2 - \frac{4}{T-1} & \text{if } t = 2, \\ 5 - \frac{1}{T-1} & \text{if } t \in \{1, T\}, , \\ 6 & \text{otherwise .} \end{cases}$$

$$\sum_{t=1}^{T/2} (x_t - \bar{x}_1)^2 = 2\sum_{t=1}^{T/2} x_t^2 - T\bar{x}_1^2 = 5 - \frac{2}{T}, \qquad \sum_{t=T/2+1}^{T} (x_t - \bar{x}_2)^2 = 1 - \frac{2}{T},$$

so we get

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_{\text{PJK}}] - \theta &= \frac{T\sigma^2}{\sum_{t=1}^T x_t^2} - \sigma^2 \frac{(T-1)}{T} \sum_{t=1}^T \frac{1}{\sum_{s\neq t}(x_s - \bar{x}_{-t})^2} \\
&= \sigma^2 \frac{T}{6} - \sigma^2 \frac{T-1}{T} \left(\frac{2}{5 - \frac{1}{T-1}} + \frac{1}{2 - \frac{4}{T-1}} + \frac{T-3}{6} \right) \\
&= \sigma^2 \left(\frac{2}{3} - \frac{4}{6T} - \frac{T-1}{T} \frac{2}{5 - \frac{1}{T-1}} - \frac{T-1}{T} \frac{1}{2 - \frac{4}{T-1}} \right) \\
&= -\frac{7}{30}\sigma^2 + O\left(\frac{1}{T}\right) \\
\text{and} \quad \mathbb{E}[\hat{\theta}_{\text{SPJK}}] - \theta &= \frac{2\sigma^2}{\sum_{t=1}^T x_t^2} - \frac{\sigma^2}{2\sum_{t=1}^{T/2}(x_t - \bar{x}_1)^2} + \frac{\sigma^2}{2\sum_{t=T/2+1}^{T}(x_t - \bar{x}_2)^2} \\
&= \sigma^2 \left(\frac{1}{3} - \frac{1}{10 - \frac{4}{T}} - \frac{1}{2 - \frac{4}{T}} \right) = -\frac{8}{30}\sigma^2 + O\left(\frac{1}{T}\right),
\end{aligned}$$

where the biases are increasing functions of T.

C4 Distribution Theory

This appendix provides restatements and proofs of Propositions 1 and 2 and Theorem 1. The proofs of the last two results relies on an auxiliary lemma which extends a central limit theorem given in Sølvsten (2017).

C4.1 The low rank case

Proposition C4.1. *If Assumption* 1 *holds,* $\max_i w_i' w_i = o(1)$*, and* r *is fixed, then*

$$\hat{\theta} = \sum_{\ell=1}^r \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell] \right) + o_p(\mathbb{V}[\hat{\theta}]^{1/2}) \quad \textit{and} \quad \mathbb{V}[\hat{b}]^{-1/2}(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, I_r)$$

where $b = Q' S_{xx}^{1/2} \beta$*, and* $\mathbb{V}[\hat{b}] = \sum_{i=1}^n w_i w_i' \sigma_i^2$*.*

*Proof of Proposition C*4.1. The proof has two steps: First, we write $\hat{\theta}$ as $\sum_{\ell=1}^r \lambda_\ell \left(\hat{b}_\ell^2 - \mathbb{V}[\hat{b}_\ell] \right)$ plus an approximation error which is of smaller order than $\mathbb{V}[\hat{\theta}]$. Second, we use Lyapounov'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

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

where we now show that the mean zero random 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

$$\sum_{i=1}^{n} B_{ii}(\hat{\sigma}_i^2 - \sigma_i^2) = \sum_{i=1}^{n} B_{ii} \sum_{\ell=1}^{n} M_{ii}^{-1} x_i' \beta M_{i\ell} \varepsilon_\ell \tag{C1}$$

$$+ \sum_{i=1}^{n} B_{ii}(\varepsilon_i^2 - \sigma_i^2) \tag{C2}$$

$$+ \sum_{i=1}^{n} B_{ii} \sum_{\ell \neq i} M_{ii}^{-1} M_{i\ell} \varepsilon_i \varepsilon_\ell. \tag{C3}$$

The variances of these three terms are

$$\text{(C1)}: \quad \sum_{\ell=1}^{n} \sigma_\ell^2 \left(\sum_{i=1}^{n} M_{i\ell} B_{ii} M_{ii}^{-1} x_i' \beta \right)^2 \leq \max_i \sigma_i^2 \sum_{i=1}^{n} B_{ii}^2 M_{ii}^{-2} (x_i' \beta)^2$$

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

$$\text{(C2)}: \quad \sum_{i=1}^{n} B_{ii}^2 \mathbb{V}[\varepsilon_i^2] \leq \max_i \mathbb{E}[\varepsilon_i^4] \times \sum_{i=1}^{n} B_{ii}^2,$$

$$\text{(C3)}: \quad \sum_{i=1}^{n} \sum_{\ell \neq i} \left(B_{ii}^2 M_{ii}^{-2} + B_{ii} M_{ii}^{-1} B_\ell M_{\ell\ell}^{-1} \right) M_{i\ell}^2 \sigma_i^2 \sigma_j^2 \leq 2 \max_i \sigma_i^4 M_{ii}^{-2} \times \sum_{i=1}^{n} B_{ii}^2.$$

Furthermore, we have that

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

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

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 $\mathbb{V}[v'\hat{b}]^{-1/2} v'(\hat{b} - b) \xrightarrow{d} N(0,1)$ if

$$\mathbb{V}[v'\hat{b}]^{-2} \sum_{i=1}^{n} \mathbb{E}[\varepsilon_i^4](v'Q'S_{xx}^{-1/2}x_i)^4 = \mathbb{V}[v'\tilde{\beta}]^{-2} \sum_{i=1}^{n} \mathbb{E}[\varepsilon_i^4](v'w_i)^4 = o(1). \tag{C4}$$

We have that $\max_i w_i'w_i = o(1)$ implies (C4) since $\max_i (v'w_i)^2 \leq \max_i w_i'w_i$ and

$$\sum_{i=1}^{n} (v'w_i)^2 = 1, \qquad \mathbb{V}[v'\tilde{\beta}]^{-1} \leq \max_i \sigma_i^{-2} = O(1), \qquad \max_i \mathbb{E}[\varepsilon_i^4] = O(1),$$

by definition of w_i and Assumption 1. □

A central limit theorem

The proofs of Proposition 2 and Theorem 1 is 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 $v_{n,i}$, $\sigma_{n,i}^2$, $\dot{w}_{n,i}$ and W_n. Define

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

Lemma C4.1. *If* $\max_i \mathbb{E}[v_i^4] + \sigma_i^{-2} = O(1)$,

$$(i) \ \max_i \dot{w}_i^2 = o(1), \qquad (ii) \ \max_\ell \lambda_\ell^2(W) = o(1),$$

then $(\mathcal{S}_n, \mathcal{U}_n)' \xrightarrow{d} \mathcal{N}(0, I_2)$.

This lemma extends the main result of Appendix S.2 in Sølvsten (2017) to allow for $\{v_i\}_i$ to be an array of non-identically distributed variables and presents the conclusion in a way that is tailored to the application in this paper. The proof requires no substantially new ideas compared to Sølvsten (2017), but we give it at the end of this section for completeness.

Proposition C4.2. *If*

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

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

C4.2 The high rank case

Proof of Proposition 2. The proof involves two steps: First, we decompose $\hat{\theta}$ into a weighted sum of two terms of the type described in Lemma $C4.1$. Second, we use Lemma $C4.1$ 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

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

where these two terms are uncorrelated and have variances

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

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

$$\mathcal{S}_n=V_{\mathcal{S}}^{-1/2}\sum_{i=1}^n\left(2\tilde{x}_i'\beta-\check{x}_i'\beta\right)\varepsilon_i,\qquad \mathcal{U}_n=V_{\mathcal{U}}^{-1/2}\sum_{i=1}^n\sum_{\ell\neq i}C_{i\ell}\varepsilon_i\varepsilon_\ell,$$
$$\omega_1=\sqrt{V_{\mathcal{S}}/\mathbb{V}[\hat{\theta}]},\qquad \omega_2=\sqrt{V_{\mathcal{U}}/\mathbb{V}[\hat{\theta}]}.$$

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 it follows from Result $C4.2$ in the next section and Proposition 2(ii) that $\hat{\theta}$ is asymptotically normal. Thus we consider the case where the limit of ω_1 is nonzero.

In the notation of Lemma $C4.1$ we have

$$\dot{w}_i = \frac{\left(2\tilde{x}_i'\beta - \check{x}_i'\beta\right)}{V_{\mathcal{S}}^{1/2}} \quad \text{and} \quad W_{i\ell} = \frac{C_{i\ell}}{V_{\mathcal{U}}^{1/2}}.$$

For Lemma $C4.1$(i) we have

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

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

For Lemma $C4.1$(ii) we show instead that $\text{trace}(W^4) = o(1)$. It can be shown that for all n, $\text{trace}(C^4) \leq c_U \cdot \text{trace}(B^4) = c_U \cdot \text{trace}(\tilde{A}^4) \leq c_U \lambda_1^2 \cdot \text{trace}(\tilde{A}^2)$ and $V_{\mathcal{U}} \geq c_L \min_i \sigma_i^4 \cdot \text{trace}(\tilde{A})$, 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

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

where the last equality follows from Proposition 2(ii). □

C4.3 The general case

Theorem C4.1. *If* $\max_i \mathsf{w}_{iq}'\mathsf{w}_{iq} = o(1)$, $\mathbb{V}[\hat{\theta}_q]^{-1} \max_i \left((\tilde{x}_{iq}'\beta)^2 + (\check{x}_{iq}'\beta)^2\right) = o(1)$, *and Assumptions* 1 *and* 2 *holds, then*

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

where

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

$$\mathbb{V}[(\hat{\mathsf{b}}_q', \hat{\theta}_q)'] = \sum_{i=1}^{n} \begin{bmatrix} \mathsf{w}_{iq}\mathsf{w}_{iq}'\sigma_i^2 & 2\mathsf{w}_{iq}\left(\sum_{\ell \neq i} C_{i\ell q} x_\ell'\beta\right)\sigma_i^2 \\ 2\mathsf{w}_{iq}'\left(\sum_{\ell \neq i} C_{i\ell q} x_\ell'\beta\right)\sigma_i^2 & 4\left(\sum_{\ell \neq i} C_{i\ell q} x_\ell'\beta\right)^2 \sigma_i^2 + 2\sum_{\ell \neq i} C_{i\ell q}^2 \sigma_i^2 \sigma_\ell^2 \end{bmatrix},$$

$C_{i\ell q} = B_{i\ell q} - 2^{-1} M_{i\ell} \left(M_{ii}^{-1} B_{iiq} + M_{\ell\ell}^{-1} B_{\ell\ell q} \right)$, *and*

$$B_{i\ell q} = x_i' S_{xx}^{-1/2} \tilde{A}_q S_{xx}^{-1/2} x_\ell \qquad \text{for } \tilde{A}_q = \tilde{A} - \sum_{\ell=1}^{q} \lambda_\ell q_\ell q_\ell',$$

$$\tilde{x}_{iq} = \sum_{\ell=1}^{n} B_{i\ell q} x_\ell, \qquad \breve{x}_{iq} = \sum_{\ell=1}^{n} M_{i\ell} M_{\ell\ell}^{-1} B_{\ell\ell q} x_\ell,$$

*Proof of Theorem C*4.1. 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 C4.1. Second, we use Lemma C4.1 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

We have that

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

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

since

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

and

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

where $B_{ii,-q} = B_{ii} - B_{iiq}$ and it follows from $\max_i \mathsf{w}_{iq}' \mathsf{w}_{iq} = o(1)$ and the calculations in the proof of Proposition 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

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

where

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

$$V_{\mathcal{S}} = \sum_{i=1}^{n} \left(2\tilde{x}_{iq}'\beta - \check{x}_{iq}'\beta\right)^2 \sigma_i^2, \qquad V_{\mathcal{U}} = 2 \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell q}^2 \sigma_i^2 \sigma_\ell^2,$$

$$\omega_1 = \sqrt{V_{\mathcal{S}} / \mathbb{V}[\hat{\theta}_q]}, \qquad \omega_2 = \sqrt{V_{\mathcal{U}} / \mathbb{V}[\hat{\theta}_q]},$$

and $\mathcal{U}_n$ is uncorrelated with both $\mathcal{S}_n$ and $\hat{\mathsf{b}}_q$.

Asymptotic Normality

As in the proof of Proposition 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 $C4.1$ 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 $C4.1$ to show that

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

for any non-random $(v', u)' \in \mathbb{R}^{q+1}$ with $v'v + u^2 = 1$. In the notation of Lemma $C4.1$ we have

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

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 1(i), Theorem 1(ii), and ω_1 being bounded away from zero.

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

$$\operatorname{trace}(C_2^4) \leq c\operatorname{trace}(B_2^4) \leq c\lambda_{q+1}^2 \sum_{\ell=q+1}^{r} \lambda_\ell^2$$

$$V_{\mathcal{U}}^2 \geq \omega_2^{-4} \min_i \sigma_i^8 \operatorname{trace}(\tilde{A}^2)^2,$$

so it follows from Assumptions 1 and 2 that $\operatorname{trace}(W^4) = o(1)$. □

Proof of a central limit theorem

The proof of Lemma $C4.1$ uses the notation and verifies the conditions of Lemmas S2.1 and S2.2 in Sølvsten (2017) referred to as SS2.1 and SS2.2, respectively. First, we show marginal convergence in distribution of $\mathcal{S}_n$ and $\mathcal{U}_n$. Then, we show joint convergence in distribution of $\mathcal{S}_n$ and $\mathcal{U}_n$. Let $V_n = (v_1, \dots, v_n)$ where $\{v_i\}_i$ are as in the setup of Lemma $C4.1$.

Before starting we note that $\max_i \sigma_i^{-2} = O(1)$ and $2\sum_{i=1}^n \sum_{\ell \neq i} W_{i\ell}^2 \sigma_i^2 \sigma_\ell^2 = 1$ implies that $\operatorname{trace}(W^2) = \sum_{i=1}^n \sum_{\ell \neq i} W_{i\ell}^2 = O(1)$ and therefore that

$$\max_\ell \lambda_\ell^2(W) = o(1) \Leftrightarrow \operatorname{trace}(W^4) = o(1).$$

Marginal Distributions

Result C4.1. $\max_i \mathbb{E}[v_i^4] + \sigma_i^{-2} = O(1)$, $\sum_{i=1}^n \dot{w}_i^2 \sigma_i^2 = 1$, *and Lemma* $C4.1$*(i) implies that* $\mathcal{S}_n \xrightarrow{d} \mathcal{N}(0,1)$.

In the notation of SS2.1 we have,

$$\Delta_i^0 \mathcal{S}_n = \dot{w}_i v_i \quad \text{and} \quad E[T_n \,|\, V_n] = 1 + \tfrac{1}{2} \sum_{i=1}^n \dot{w}_i^2 (v_i^2 - \sigma_i^2),$$

and it follows from $\max_i \mathbb{E}[v_i^4] + \sigma_i^{-2} = O(1)$, $\sum_{i=1}^n \dot{w}_i^2 \sigma_i^2 = 1$, and Lemma $C4.1$(i) that

$$E[T_n \,|\, V_n] \xrightarrow{\mathcal{L}^1} 1, \quad \sum_{i=1}^n \mathbb{E}[(\Delta_i^0 \mathcal{S}_n)^2] = 1, \qquad \sum_{i=1}^n \mathbb{E}[(\Delta_i^0 \mathcal{S}_n)^4] \leq \max_i \frac{\mathbb{E}[v_i^4]}{\sigma_i^2} \dot{w}_i^2 = o(1),$$

so Result $C4.1$ follows from SS2.1.

Result C4.2. $\max_i \mathbb{E}[v_i^4] + \sigma_i^{-2} = O(1)$, $2\sum_{i=1}^n \sum_{\ell \neq i} W_{n,i\ell}^2 \sigma_{n,i}^2 \sigma_{n,\ell}^2 = 1$, *and Lemma* $C4.1$*(ii) implies that* $\mathcal{U}_n \xrightarrow{d} \mathcal{N}(0,1)$.

In the notation of SS2.1 we have,

$$\Delta_i^0 \mathcal{U}_n = 2 v_i \sum_{\ell \neq i} W_{i\ell} v_\ell \quad \text{and} \quad E[T_n \,|\, V_n] = \sum_{i=1}^n \sum_{\ell \neq i} \sum_{k \neq i} (v_i + \sigma_i^2) W_{i\ell} W_{ik} v_\ell v_k,$$

and

$$\sum_{i=1}^n \mathbb{E}[(\Delta_i^0 \mathcal{U}_n)^2] = 2, \qquad \sum_{i=1}^n \mathbb{E}[(\Delta_i^0 \mathcal{U}_n)^4] \leq 2^5 \max_i \mathbb{E}[v_i^4]^2 \max_i \sigma_i^{-4} \max_i \sum_{\ell \neq i} W_{i\ell}^2,$$

where $\max_i \sum_{\ell \neq i} W_{i\ell}^2 \leq \sqrt{\text{trace}(W^4)} = o(1)$. Now, split $E[T_n \,|\, V_n] - 1$ into three terms

$$a_n = \sum_{i=1}^{n} \sum_{\ell \neq i} \sigma_i^2 W_{i\ell}^2 (v_\ell + v_\ell^2 - \sigma_\ell^2)$$
$$b_n = 2 \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{k \neq i,\ell} \sigma_k^2 W_{\ell k} W_{ik} v_i v_\ell + \sum_{i=1}^{n} \sum_{\ell \neq i} W_{i\ell}^2 v_i (v_\ell^2 - \sigma_\ell^2)$$
$$c_n = \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{k \neq i,\ell} W_{i\ell} W_{ik} (v_i^2 - \sigma_i^2) v_\ell v_k.$$

Interlude: Convergence in $\mathcal{L}^1$

a_n, b_n, and c_n are a linear sum, a quadratic sum, and a cubic sum. We will need to treat similar sums later, so we record some simple sufficient conditions for their convergence. For brevity, let $\sum_{i \neq \ell}^{n} = \sum_{i=1}^{n} \sum_{\ell \neq i}$, and $\sum_{i \neq \ell \neq k}^{n} = \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{k \neq i,\ell}$, etc. We use the notation $u_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4}) \in \mathbb{R}^4$ to denote independent random vectors in order that the result applies to combinations of v_i and $v_i^2 - \sigma_i^2$ as in a_n, b_n, and c_n above. For the inferential results we will also treat quartic sums, so we provide the sufficient conditions here.

Result C4.3. *Let* $S_{n1} = \sum_{i=1}^{n} \omega_i v_{i1}$, $S_{n2} = \sum_{i \neq \ell}^{n} \omega_{i\ell} v_{i1} v_{\ell 2}$, $S_{n3} = \sum_{i \neq \ell \neq k}^{n} \omega_{i\ell k} v_{i1} v_{\ell 2} v_{k3}$, *and* $S_{n4} = \sum_{i \neq \ell \neq k \neq m}^{n} \omega_{i\ell km} v_{i1} v_{\ell 2} v_{k3} v_{m4}$ *where the weights* ω_i, $\omega_{i\ell}$, $\omega_{i\ell k}$, *and* $\omega_{i\ell km}$ *are non-random. Suppose that* $\mathbb{E}[u_i] = 0$, $\max_i \mathbb{E}[u_i' u_i] = O(1)$.

1. *If* $\sum_{i=1}^{n} \omega_i^2 = o(1)$, *then* $S_{n1} \xrightarrow{\mathcal{L}^1} 0$.
2. *If* $\sum_{i \neq \ell}^{n} \omega_{i\ell}^2 = o(1)$, *then* $S_{n2} \xrightarrow{\mathcal{L}^1} 0$.
3. *If* $\sum_{i \neq \ell \neq k}^{n} \omega_{i\ell k}^2 = o(1)$, *then* $S_{n3} \xrightarrow{\mathcal{L}^1} 0$.
4. *If* $\sum_{i \neq \ell \neq k \neq m}^{n} \omega_{i\ell km}^2 = o(1)$, *then* $S_{n4} \xrightarrow{\mathcal{L}^1} 0$.

Consider S_{n3}, the other results follows from the same line of reasoning. In the notation of SS2.2 we have,

$$\Delta_i^0 S_{n3} = v_{i1} \sum_{\ell \neq i} \sum_{k \neq i,\ell} \omega_{i\ell k} v_{\ell 2} v_{k3} + v_{i2} \sum_{\ell \neq i} \sum_{k \neq i,\ell} \omega_{\ell i k} v_{\ell 1} v_{k3} + v_{i3} \sum_{\ell \neq i} \sum_{k \neq i,\ell} \omega_{\ell k i} v_{\ell 1} v_{k2}.$$

Focusing on the first term we have,

$$\sum_{i=1}^{n}\mathbb{E}\left[\left(v_{i1}\sum_{\ell\neq i}\sum_{k\neq i,\ell}\omega_{i\ell k}v_{\ell 2}v_{k3}\right)^2\right] \leq \max_i \mathbb{E}[u_i'u_i]^3\sum_{i\neq\ell\neq k}^{n}\left(\omega_{i\ell k}^2+\omega_{i\ell k}\omega_{ik\ell}\right)$$
$$\leq 2\max_i \mathbb{E}[u_i'u_i]^3\sum_{i\neq\ell\neq k}^{n}\omega_{i\ell k}^2,$$

so the results follows from SS2.2, $\sum_{i\neq\ell\neq k}^{n}\omega_{i\ell k}^2 = o(1)$, and the observation that the last bound also applies to the other two terms in $\Delta_i^0 S_{n3}$.

Marginal Distributions, continued

To see how $a_n \xrightarrow{\mathcal{L}^1} 0$, $b_n \xrightarrow{\mathcal{L}^1} 0$ and $c_n \xrightarrow{\mathcal{L}^1} 0$ follows from Result $C4.3$, let $\bar{W}_{i\ell} = \sum_{k=1}^{n} W_{ik}W_{k\ell}$ and note that $\text{trace}(W^4) = \sum_{i=1}^{n}\sum_{\ell=1}^{n}\bar{W}_{i\ell}^2$. We have

$$\sum_{i=1}^{n}\left(\sum_{\ell\neq i}\sigma_\ell^2 W_{i\ell}^2\right)^2 \leq \max_i\sigma_i^4\sum_{i=1}^{n}\bar{W}_{ii}^2.$$
$$\sum_{i=1}^{n}\sum_{\ell\neq i}\left(\sum_{k\neq i,\ell}\sigma_k^2 W_{\ell k}W_{ik}\right)^2 \leq \max_i\sigma_i^4\sum_{i=1}^{n}\sum_{\ell=1}^{n}\bar{W}_{i\ell}^2$$
$$\sum_{i=1}^{n}\sum_{\ell\neq i}W_{i\ell}^4 = O\left(\max_{i,\ell}W_{i\ell}^2\right)$$
$$\sum_{i=1}^{n}\sum_{\ell\neq i}\sum_{k\neq i,\ell}W_{i\ell}^2W_{ik}^2 = O\left(\max_i\sum_{\ell\neq i}W_{i\ell}^2\right),$$

all of which are $o(1)$ as $\text{trace}(W^4) = o(1)$.

Joint Distribution

Let $(u_1, u_2)' \in R^2$ be given and non-random with $u_1^2 + u_2^2 = 1$. Define $\mathcal{W}_n = u_1\mathcal{S}_n + u_2\mathcal{U}_n$. Lemma $C4.1$ follows if we show that $\mathcal{W}_n \xrightarrow{d} \mathcal{N}(0,1)$. In the notation of SS2.1 we have,

$$\Delta_i^0\mathcal{W}_n = u_1\dot{w}_i v_i + u_2 2v_i\sum_{\ell\neq i}W_{i\ell}v_\ell$$

and

$$\mathbb{E}[T_n \,|\, V_n] = u_1^2 \left(1 + \tfrac{1}{2}\sum_{i=1}^n \dot{w}_i^2 (v_i^2 - \sigma_i^2)\right) + u_2^2 \sum_{i=1}^n \sum_{\ell \neq i} \sum_{k \neq i} (v_i + \sigma_i^2) W_{i\ell} W_{ik} v_\ell v_k$$
$$+ u_1 u_2 3 \sum_{i=1}^n \sum_{\ell \neq i} (v_i^2 + \sigma_i^2) \dot{w}_i W_{i\ell} v_j.$$

The proofs of Result $C4.1$ and Result $C4.2$ showed that

$$\sum_{i=1}^n \mathbb{E}[(\Delta_i^0 \mathcal{W}_n)^2] = O(1), \ \sum_{i=1}^n \mathbb{E}[(\Delta_i^0 \mathcal{W}_n)^4] = o(1)$$

and that the first two terms of $\mathbb{E}[T_n \,|\, V_n]$ converge to $u_1^2 + u_2^2 = 1$. Thus the lemma follows if we show that the "conditional covariance"

$$3 \sum_{i=1}^n \sum_{\ell \neq i} (v_i^2 + \sigma_i^2) \dot{w}_i W_{i\ell} v_j$$

converges to 0 in $\mathcal{L}^1$. This conditional covariance involves a linear and a quadratic sum so

$$\sum_{i=1}^n \left(\sum_{\ell \neq i} \sigma_\ell^2 w_\ell W_{i\ell} \right)^2 \leq \max_i \sigma_i^4 \max_\ell \lambda_\ell^2(W) \sum_{i=1}^n \dot{w}_i^2 = O(\max_\ell \lambda_\ell^2(W))$$
$$\sum_{i=1}^n \sum_{\ell \neq i} \dot{w}_i^2 W_{i\ell}^2 \leq \sum_{i=1}^n \sum_{\ell \neq i} W_{i\ell}^2 \max_i \dot{w}_i^2 = O(\max_i \dot{w}_i^2)$$

ends the proof.

C5 Inference

C5.1 Asymptotic Variance Estimation

Lemma C5.1. *If the conditions of Proposition* 1 *holds, then* $\mathbb{V}[\hat{b}]^{-1} \hat{\mathbb{V}}[\hat{b}] \xrightarrow{p} I_r$.

Proof of Lemma $C5.1$. It suffices to show that

$$\delta(v) := \frac{\hat{\mathbb{V}}[v'\hat{b}] - \mathbb{V}[v'\hat{b}]}{\mathbb{V}[v'\hat{b}]} = o_p(1)$$

for all nonrandom $v \in \mathbb{R}^r$ with $v'v = 1$.

Let $v \in \mathbb{R}^r$ be nonrandom with $v'v = 1$. As in the first step in the proof of Proposition 1 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 $w_i(v) = \frac{(v'w_i)^2}{\sum_{i=1}^n \sigma_i^2 (v'w_i)^2}$. But this follows from

$$\sum_{i=1}^n w_i(v)^4 \leq \max_i \sigma_i^{-4} \max_i w_i' w_i = o(1)$$

by Proposition 1(i), $v'v = 1$, and $\sum_{i=1}^n w_i w_i' = I_r$. □

Lemma C5.2. *If the conditions of Proposition* 2 *and Assumption* 3 *holds, then* $\hat{\mathbb{V}}[\hat{\theta}]/\mathbb{V}[\hat{\theta}] \xrightarrow{p} 1$.

*Proof of Lemma C*5.2. It suffices to show that

$$\delta := \frac{\hat{\mathbb{V}}[\hat{\theta}] - \mathbb{V}[\hat{\theta}]}{\mathbb{V}[\hat{\theta}]} = o(1).$$

where we have that

$$\delta = 4\mathbb{V}[\hat{\theta}]^{-1} \sum_{i=1}^n \left(\sum_{\ell \neq i} C_{i\ell} x_\ell' \beta \right)^2 (\tilde{\sigma}_i^2 - \sigma_i^2) \tag{C5}$$

$$+ 4\mathbb{V}[\hat{\theta}]^{-1} \sum_{i=1}^n \left(\sum_{\ell \neq i} C_{i\ell} \varepsilon_\ell \right)^2 \tilde{\sigma}_i^2 - 2\mathbb{V}[\hat{\theta}]^{-1} \sum_{i=1}^n \sum_{\ell \neq i} C_{i\ell}^2 \left(\tilde{\sigma}_i^2 \tilde{\sigma}_\ell^2 + \sigma_i^2 \sigma_\ell^2 \right) \tag{C6}$$

$$+ 8\mathbb{V}[\hat{\theta}]^{-1} \sum_{i=1}^n \left(\sum_{\ell \neq i} C_{i\ell} x_\ell' \beta \right) \left(\sum_{\ell \neq i} C_{i\ell} \varepsilon_\ell \right) \tilde{\sigma}_i^2. \tag{C7}$$

We proceed by showing that both the mean and variance of δ converge to zero and for this we rely on the representation:

$$\tilde{\sigma}^2(\omega) = \sum_{i=1}^n k_i(\omega) \hat{\sigma}_i^2 = \sum_{i=1}^n k_i(\omega) \varepsilon_i^2 + \sum_{\ell=1}^n \left(\sum_{i=1}^n k_i(\omega) M_{ii}^{-1} x_i' \beta \right) M_{i\ell} \varepsilon_\ell$$
$$+ \sum_{i=1}^n \sum_{\ell \neq i} k_i(\omega) M_{ii}^{-1} M_{i\ell} \varepsilon_i \varepsilon_\ell \tag{C8}$$

from which it follows that

$$\left|\mathbb{E}\left[\tilde{\sigma}^2(\omega)\right] - \sigma^2(\omega)\right| = \left|\sum_{i=1}^{n} k_i(\omega)(\sigma^2(\omega_i) - \sigma^2(\omega))\right|$$

$$\leq c\sum_{i=1}^{n} |k_i(\omega)| \|\omega_i - \omega\|_M \tag{C9}$$

$$\mathbb{V}[\tilde{\sigma}^2(\omega)] = O\left(\sum_{i=1}^{n} k_i(\omega)^2\right) \tag{C10}$$

where the last equality follows from the calculations applied to (C1)–(C3) (with k_i replacing B_{ii}).

From (C9) we find that the mean of (C5) is of order

$$\max_i \sum_{\ell=1}^{n} |k_\ell(\omega_i)| \|\omega_i - \omega_\ell\|_M = o(1)$$

and from (C8) we find that the variance of (C5) is of order

$$\max_i \sum_{\ell=1}^{n} k_\ell(\omega_i)^2 = o(1).$$

Similarly, we find from (C9) and (C10) that the mean of (C6) is of order

$$\max_i \left|\mathbb{E}\left[\tilde{\sigma}^2(\omega_i)\right] - \sigma^2(\omega_i)\right| + \mathbb{V}[\tilde{\sigma}^2(\omega_i)] + \left(\sum_{\ell=1}^{n} k_\ell(\omega_i)^2\right)^{1/2} = o(1),$$

and that the mean of (C7) is of order

$$\max_i \left(\sum_{i=1}^{n} k_\ell(\omega_i)^2\right)^{1/2} = o(1).$$

Now, the demeaned versions of (C6) and (C7) involve linear, quadratic, cubic, and quartic sums. We have already treated versions of linear, quadratic and cubic sums in detail in the proof of Lemma $C4.1$. Thus, we report here the calculations for the two quartic terms stemming from (C6) (details for the remaining terms can be provided upon request):

$$\sum_{i\neq\ell\neq k\neq m}^{n} \omega_{i\ell km}\varepsilon_i\varepsilon_\ell\varepsilon_k\varepsilon_m$$

where $\omega_{i\ell km}$ is either

$$4\mathbb{V}[\hat{\theta}]^{-1}\sum_{j=1}^{n} C_{ji}C_{j\ell}k_k(\omega_j)M_{kk}^{-1}M_{km}$$

or

$$2\mathbb{V}[\hat{\theta}]^{-1}\sum_{j=1}^{n}\sum_{j'=1}^{n} C_{jj'}^2 k_i(\omega_j)M_{ii}^{-1}M_{i\ell}k_k(\omega_{j'})M_{kk}^{-1}M_{km}.$$

Some calculations yield that $\sum_{i\neq\ell\neq k\neq m}^{n}\omega_{i\ell km}^2$ is of order

$$\frac{\text{trace}(C^4)}{\mathbb{V}[\hat{\theta}]}\max_i\sum_{\ell=1}^{n}k_\ell(\omega_i)^2 + \left(\max_i\sum_{\ell=1}^{n}k_\ell(\omega_i)^2\right)^2 = o(1)$$

in either case. □

Lemma C5.3. *If the conditions of Theorem* 1 *and Assumption* 3 *hold, then*

$$\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}.$$

*Proof of Lemma C*5.3. The statements

$$\mathbb{V}[\hat{\mathsf{b}}_q]^{-1}\hat{\mathbb{V}}[\hat{\mathsf{b}}_q] \xrightarrow{p} I_q \quad \text{and} \quad \frac{\hat{\mathbb{V}}[\hat{\theta}_q]}{\mathbb{V}[\hat{\theta}_q]} \xrightarrow{p} 1$$

follows by applying the arguments in Lemmas *C*5.1 and *C*5.2. Thus we focus on the remaining claim that

$$\delta(v) := \frac{\hat{\mathcal{C}}[v'\hat{\mathsf{b}}_q,\hat{\theta}_q] - \mathcal{C}[v'\hat{\mathsf{b}}_q,\hat{\theta}_q]}{\mathbb{V}[v'\hat{\mathsf{b}}_q]^{1/2}\mathbb{V}[\hat{\theta}_q]^{1/2}} \xrightarrow{p} 0 \quad \text{where} \quad \hat{\mathcal{C}}[v'\hat{\mathsf{b}}_q,\hat{\theta}_q] = 2\sum_{i=1}^{n} v'\mathsf{w}_{iq}\left(\sum_{\ell\neq i}C_{i\ell q}y_\ell\right)\tilde{\sigma}_i^2$$

for all non-random $v \in \mathbb{R}^q$ with $v'v = 1$. The calculations and arguments used are repetitions of those used to handle (C5) and (C6) so it follows from (C9) and (C10) that the expectation of $\delta(v)$ is of order

$$\max_i\sum_{\ell=1}^{n}|k_\ell(\omega_i)|\|\omega_i-\omega_\ell\|_M + \max_i\left(\sum_{i=1}^{n}k_\ell(\omega_i)^2\right)^{1/2} = o(1),$$

and the variance of $\delta(v)$ can similarly be shown to be of order $o(1)$ by applying Result *C*4.3. □

C5.2 Confidence Intervals

Critical value function

For a given curvature $\kappa > 0$ and confidence level $1-\alpha$, the critical value function z_κ is the square-root of the $(1-\alpha)$'th quantile of

$$\min_{-\frac{2}{\kappa} \leq x \leq 0} \left(\sqrt{\frac{1}{\kappa^2} - \left(\frac{1}{\kappa} + x\right)^2} - \sqrt{\chi_q^2} \right)^2 + \left(x - \sqrt{\chi_1^2} \right)^2$$

where χ_q^2 and χ_1^2 are independently distributed. The critical value function at $\kappa = 0$ is the limit of z_κ as $\kappa \downarrow 0$, which is the $(1-\alpha/2)$'th quantile of a standard normal random variable. See Andrews and Mikusheva (2016) for additional details.

Curvature

The confidence interval $\hat{C}_q^\theta$ inverts hypotheses of the type $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on the value of the test statistic

$$\min_{\mathsf{b}_q, \theta_q : g(\mathsf{b}_q, \theta_q) = 0} \begin{pmatrix} \hat{\mathsf{b}}_q - \mathsf{b}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix}' \hat{\Sigma}^{-1} \begin{pmatrix} \hat{\mathsf{b}}_q - \mathsf{b}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix}$$

where $g(\mathsf{b}_q, \theta_q) = \sum_{\ell=1}^q \lambda_\ell \mathsf{b}_{q,\ell}^2 + \theta_q - \theta_0$ and $\mathsf{b}_q = (\mathsf{b}_{q,1}, \ldots, \mathsf{b}_{q,q})'$. This testing problem depends on the manifold

$$\hat{S} = \left\{ x = \hat{\Sigma}_q^{-1/2} (\mathsf{b}_q, \theta_q)' : g(\mathsf{b}_q, \theta_q) = 0 \right\}$$

for which we need an upper bound on the maximal curvature. To derive this upper bound we look at the parameterization

$$\mathbf{x}(\dot{y}) = \hat{\Sigma}_q^{-1/2} (\dot{y}_1, \ldots, \dot{y}_q, \theta_0 - \sum_{\ell=1}^q \lambda_\ell \dot{y}_\ell^2)'$$

which maps from $\mathbb{R}^q$ to $\hat{S}$, is a homeomorphism, and has a Jacobian of full rank:

$$d\mathbf{x}(\dot{y}) = \hat{\Sigma}_q^{-1/2} \begin{bmatrix} \operatorname{diag}(1, \ldots, 1) \\ -2\lambda_1 \dot{y}_1, \ldots, -2\lambda_q \dot{y}_q \end{bmatrix} := Z_y$$

$\hat{\kappa}$, the curvature of $\hat{S}$, is then given as $\hat{\kappa} = \max_{\dot{y}\in\mathbb{R}^q} \kappa_{\dot{y}}$

$$\kappa_{\dot{y}} = \sup_{u\in\mathbb{R}^q} \frac{\|(I - P_{\dot{y}})V(u \odot u)\|}{\|Z_{\dot{y}}u\|^2}$$

where $P_{\dot{y}} = Z_{\dot{y}}(Z'_{\dot{y}}Z_{\dot{y}})^{-1}Z'_{\dot{y}}$ and V is the matrix of second derivatives of $d\mathbf{x}(\dot{y})$

$$V = \hat{\Sigma}_q^{-1/2} \begin{bmatrix} 0 \\ -2\lambda_1, \ldots, -2\lambda_q \end{bmatrix}.$$

Curvature when $q = 1$

In this case the maximization over u drops out and we have

$$\hat{\kappa} = \max_{y\in\mathbb{R}} \frac{\sqrt{V'V - \frac{(v'V)^2}{v'v}}}{v'v}$$

where $v = \hat{\Sigma}_q^{-1/2}(1, -2\lambda_1\dot{y})'$ and $V = \hat{\Sigma}_q^{-1/2}(0, -2\lambda_1)$. The value $\dot{y} = -\frac{\hat{\rho}\frac{\hat{\mathbb{V}}[\hat{\theta}_q]}{\hat{\mathbb{V}}[\hat{b}_1]}}{2\lambda_1}$ is both a minimizer of $v'v$ and of $(v'V)^2$, and it therefore leads to $\hat{\kappa} = \frac{2|\lambda_1|\hat{\mathbb{V}}[\hat{b}_1]}{\hat{\mathbb{V}}[\hat{\theta}_q]^{1/2}(1-\hat{\rho}^2)^{1/2}}$.

Curvature when $q > 1$

In this case we first maximize over $\dot{y}$ and then over u. For a fixed u we want to find

$$\max_{\dot{y}\in\mathbb{R}^q} \frac{\sqrt{V'_uV_u - V'_uP_{\dot{y}}V_u}}{v'_{u,\dot{y}}v_{u,\dot{y}}}$$

where $V_u = \hat{\Sigma}_q^{-1/2}(0, -2\sum_{\ell=1}^q \lambda_\ell u_\ell^2)$, $v_{u,\dot{y}} = \hat{\Sigma}_q^{-1/2}(u', -2u'D_q\dot{y})'$ and $D_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. The value for $\dot{y}$ that solves $-2D_q\dot{y} = \hat{\mathbb{V}}[\hat{\mathsf{b}}_q]^{-1}\hat{\mathbb{C}}[\hat{\mathsf{b}}_q, \hat{\theta}_q]$ sets $P_{\dot{y}}V_u = 0$ and minimizes $v'_{u,\dot{y}}v_{u,\dot{y}}$. Thus we obtain

$$\begin{aligned}\hat{\kappa} &= \frac{2}{\left(\hat{\mathbb{V}}[\hat{\theta}_q] - \hat{\mathbb{C}}[\hat{\mathsf{b}}_q, \hat{\theta}_q]'\hat{\mathbb{V}}[\hat{\mathsf{b}}_q]^{-1}\hat{\mathbb{C}}[\hat{\mathsf{b}}_q, \hat{\theta}_q]\right)^{1/2}} \max_{u\in\mathbb{R}^q} \frac{|u'D_qu|}{u'\hat{\mathbb{V}}[\hat{\mathsf{b}}_q]^{-1}u} \\ &= \frac{2|\dot{\hat{\lambda}}_1(\hat{\mathbb{V}}[\hat{\mathsf{b}}_q]^{1/2}D_q\hat{\mathbb{V}}[\hat{\mathsf{b}}_q]^{1/2})|}{\left(\hat{\mathbb{V}}[\hat{\theta}_q] - \hat{\mathbb{C}}[\hat{\mathsf{b}}_q, \hat{\theta}_q]'\hat{\mathbb{V}}[\hat{\mathsf{b}}_q]^{-1}\hat{\mathbb{C}}[\hat{\mathsf{b}}_q, \hat{\theta}_q]\right)^{1/2}}\end{aligned}$$

where $\dot{\hat{\lambda}}_1(\cdot)$ is the eigenvalue of largest magnitude.

Closed form representation of $\hat{C}_1^\theta$

The upper end of the interval is found by noting that maximization over a linear function in θ_2 implies that the constraint must bind at the maximum, so we can reformulate the bivariate problem as a univariate problem

$$\begin{aligned}
&\max_{\dot{b}_1,\theta_2}\left\{\lambda_1\dot{b}_1^2+\theta_1:\begin{pmatrix}\hat{b}_1-\dot{b}_1\\ \hat{\theta}_1-\theta_1\end{pmatrix}'\hat{\Sigma}_1^{-1}\begin{pmatrix}\hat{b}_1-\dot{b}_1\\ \hat{\theta}_1-\theta_1\end{pmatrix}\leq z_{\hat{\kappa}}^2\right\}\\
=&\max_{\dot{b}_1}\lambda_1 b_1^2+\hat{\theta}_1-\hat{\rho}\frac{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}}(\hat{b}_1-b_{1,\pm})+\left(\hat{\mathbb{V}}[\hat{\theta}_1](1-\hat{\rho}^2)\right)^{1/2}\left(z_{\hat{\kappa}}^2-\frac{(\hat{b}_1-b)^2}{\hat{\mathbb{V}}[\hat{b}_1]}\right)^{1/2}
\end{aligned}$$

where we are implicitly enforcing the constraint on $\dot{b}_1$ that the term under the square-root is non-negative. Thus we will find a global max in $\dot{b}_1$ and note that it satisfies the constraint. The first order condition is

$$2\lambda_1 b_{1,+}+\hat{\rho}\frac{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}}+\left(\hat{\mathbb{V}}[\hat{\theta}_1](1-\hat{\rho}^2)\right)^{1/2}\frac{\frac{\hat{b}_1-b_{1,+}}{\hat{\mathbb{V}}[\hat{b}_1]}}{\left(z_{\hat{\kappa}}^2-\frac{(\hat{b}_1-b_{1,+})^2}{\hat{\mathbb{V}}[\hat{b}_1]}\right)^{1/2}}=0$$

which after some rearrangement and squaring of both sides implies that

$$\frac{(\hat{b}_1-b_{1,+})^2}{\hat{\mathbb{V}}[\hat{b}_1]}=(1-\hat{\alpha}_+)z_{\hat{\kappa}}^2.$$

All solutions, $b_{1,+}$, to this equation satisfies the non-negativity constraint since

$$\left(z_{\hat{\kappa}}^2-\frac{(\hat{b}_1-b)^2}{\hat{\mathbb{V}}[\hat{b}_1]}\right)^{1/2}=z_{\hat{\kappa}}\hat{\alpha}_+^{1/2}\geq 0.$$

Inserting this in the first order condition yields the implicit solution

$$b_+=\hat{b}_1+z_{\hat{\kappa}}\left(\hat{\mathbb{V}}[\hat{b}_1](1-\hat{a}_+)\right)^{1/2}$$

and upper bound of

$$f_+(b_{1,+})=\lambda_1 b_{1,+}^2+\hat{\theta}_1-\hat{\rho}\frac{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}}(\hat{b}_1-b_{1,+})+z_{\hat{\kappa}}\left(\hat{\mathbb{V}}[\hat{\theta}_1](1-\hat{\rho}^2)\hat{a}_+\right)^{1/2}.$$

Rearranging and squaring the first order condition yields $b_{1,+}$ as a solution to the quartic equation:

$$\frac{(\hat{b}_1 - \dot{b}_1)^2}{\hat{\mathbb{V}}[\hat{b}_1]}\left(1 + \left(\frac{\text{sgn}(\lambda_1)\hat{\kappa}\dot{b}_1}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}} + \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\right)^2\right) = \left(\frac{\text{sgn}(\lambda_1)\hat{\kappa}\dot{b}_1}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}} + \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\right)^2 z_{\hat{\kappa}}^2. \quad \text{(C11)}$$

Thus the upper end of the confidence set can be found by maximizing f_+ over the at most four real solutions to the fourth order polynomial in (C11) that are also solutions to the first and second order conditions

$$\begin{aligned} b_{1,+} &= \hat{b}_1 + z_{\hat{\kappa}}\left(\hat{\mathbb{V}}[\hat{b}_1](1-\hat{a}_+)\right)^{1/2} \\ \text{sgn}(\lambda_1)\hat{\kappa} z_{\hat{\kappa}} &\leq \hat{a}_+^{-3/2}\left(1-\hat{\rho}^2\right) \end{aligned}$$

The same set of calculations yield that the lower end of the confidence set can be found by minimizing

$$f_-(b_{1,-}) = \lambda_1 b_{1,-}^2 + \hat{\theta}_1 - \hat{\rho}\frac{\hat{\mathbb{V}}[\hat{\theta}_1]^{1/2}}{\hat{\mathbb{V}}[\hat{b}_1]^{1/2}}(\hat{b}_1 - b_{1,-}) - z_{\hat{\kappa}}\left(\hat{\mathbb{V}}[\hat{\theta}_1](1-\hat{\rho}^2)\hat{a}_-\right)^{1/2}$$

over the real solutions to (C11) that are also solutions to the first and second order conditions

$$\begin{aligned} b_{1,-} &= \hat{b}_1 - z_{\hat{\kappa}}\left(\hat{\mathbb{V}}[\hat{b}_1](1-\hat{a}_-)\right)^{1/2} \\ \text{sgn}(\lambda_1)\hat{\kappa} z_{\hat{\kappa}} &\geq -\hat{a}_-^{-3/2}\left(1-\hat{\rho}^2\right). \end{aligned}$$

Validity

Lemma C5.4. *If $\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}$ and the conditions of Theorem 1 hold, then*

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta \in \hat{C}_q^{\theta}\right) \geq 1-\alpha.$$

Proof. The following two conditions are the inputs to the proof of Theorem 2 in Andrews and Mikusheva (2016), from which it follows that

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta \in \hat{C}_q^{\theta}\right) = \liminf_{n\to\infty} \mathbb{P}\left(\min_{(\dot{\mathsf{b}}_q', \theta_q)' \in \mathcal{B}} \begin{pmatrix} \hat{\mathsf{b}}_q - \dot{\mathsf{b}}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix}' \hat{\Sigma}_q^{-1} \begin{pmatrix} \hat{\mathsf{b}}_q - \dot{\mathsf{b}}_q \\ \hat{\theta}_q - \theta_q \end{pmatrix} \leq z_{\hat{\kappa}}^2\right) \geq 1-\alpha$$

where $\mathcal{B} = \left\{(\dot{\mathsf{b}}_q', \theta_q)' : \sum_{\ell=1}^q \lambda_\ell \dot{\mathsf{b}}_{q,\ell}^2 + \theta_q - \theta = 0\right\}$ and the last inequality follows from Theorem 2 in Andrews and Mikusheva (2016).

Condition (i):

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

which follows from Theorem 1 and $\Sigma_q^{-1}\hat{\Sigma}_q \xrightarrow{p} I_{q+1}$.

Condition (ii): The conditions of Lemma 1 in Andrews and Mikusheva (2016) are satisfied. To verify this, take the manifold

$$\tilde{S} = \left\{\dot{x} \in \mathbb{R}^{q+1} : \tilde{g}(x) = 0\right\}$$

for

$$\tilde{g}(\dot{x}) = \dot{x}'\hat{\Sigma}_q^{1/2}\begin{bmatrix} D_q & 0 \\ 0 & 0 \end{bmatrix}\hat{\Sigma}_q^{1/2}\dot{x} + (2\mathbb{E}[\hat{\mathsf{b}}_q]', 1)\begin{bmatrix} D_q & 0 \\ 0 & 1 \end{bmatrix}\hat{\Sigma}_q^{1/2}\dot{x}.$$

The curvature of $\tilde{S}$ is the same as that of $\hat{S}$, $\tilde{g}(0) = 0$, and $\tilde{g}$ is continuously differentiable with a Jacobian of rank 1. These are the conditions of Lemma 1 in Andrews and Mikusheva (2016). This finishes the proof.

□

C6 Verifying Conditions

Example 1. The only non-immediate conclusions are that:

$$\begin{aligned}
\mathbb{V}[\hat{\theta}]^{-1}\max_i(\tilde{x}_i'\beta)^2 &= O\left(\frac{\max_i(x_i'\beta)^2/n^2}{\min_i \sigma_i^2 \text{trace}(\tilde{A}^2)}\right) = O\left(\frac{\max_i(x_i'\beta)^2}{r}\right) \\
\mathbb{V}[\hat{\theta}]^{-1}\max_i(\check{x}_i'\beta)^2 &= O\left(\frac{\max_{i,j} M_{jj}^{-2}\left(P_{jj} - \frac{1}{n}\right)^2(x_j'\beta)^2\left(\sum_{\ell=1}^n |M_{i\ell}|\right)^2/n^2}{\min_i \sigma_i^2 \text{trace}(\tilde{A}^2)}\right) \\
&= O\left(\frac{\max_{i,j}(x_j'\beta)^2\left(\sum_{\ell=1}^n |M_{i\ell}|\right)^2}{r}\right).
\end{aligned}$$

Example 2. We first derive the representations of $\hat{\sigma}_\alpha^2$ given in section 2. When there are no common regressors, the representation in (4) follows from $B_{ii} = \frac{1}{nT_{g(i)}}\left(1 - \frac{T_{g(i)}}{n}\right)$ and

$$\hat{\sigma}_g^2 = \frac{1}{T_g}\sum_{t=1}^{T_g} y_{gt}\left(y_{gt} - \frac{1}{T_g - 1}\sum_{s\neq t} y_{gs}\right) = \frac{1}{T_g}\sum_{i:g(i)=g}\hat{\sigma}_i^2$$

which yields that

$$\sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 = \frac{1}{n}\sum_{g=1}^{N}\left(1-\frac{T_g}{n}\right)\hat{\sigma}_g^2.$$

With common regressors, it follows from the formula for block inversion of matrices that

$$\begin{aligned}\tilde{X}' = AS_{xx}^{-1}\begin{bmatrix}D'\\X'\end{bmatrix} &= \frac{1}{n}\begin{bmatrix}\left(D'-\bar{d}\mathbf{1}_n'\right)\left(I-X\left(X'(I-P_D)X'\right)^{-1}X'(I-P_D)\right)\\0\end{bmatrix}\\ &= \frac{1}{n}\begin{bmatrix}D'-\bar{d}\mathbf{1}_n'-\hat{\Gamma}'X'(I-P_D)\\0\end{bmatrix}\end{aligned}$$

where $D=(d_1,\ldots,d_n)'$, $X=(x_{g(1)t(1)},\ldots,x_{g(n)t(n)})'$, $P_D=DS_{dd}^{-1}D'$, $\mathbf{1}_n=(1,\ldots,1)'$, and $S_{dd}=D'D$. Thus it follows that

$$\tilde{x}_i = \frac{1}{n}\begin{pmatrix}d_i-\bar{d}-\hat{\Gamma}'(x_{g(i)t(i)}-\bar{x}_{g(i)})\\0\end{pmatrix}.$$

The no common regressors claims are immediate. With common regressors we have

$$P_{i\ell} = T_{g(i)}^{-1}\mathbf{1}_{\{g(i)=g(\ell)\}} + n^{-1}(x_{g(i)t(i)}-\bar{x}_{g(i)})'W^{-1}(x_{g(\ell)t(\ell)}-\bar{x}_{g(\ell)}) = T_{g(i)}^{-1}\mathbf{1}_{\{i=\ell\}}+O(n^{-1})$$

where $W=\frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T}(x_{gt}-\bar{x}_g)(x_{gt}-\bar{x}_g)'$ so $P_{ii}\le C<1$ in large samples. The eigenvalues of $\tilde{A}$ are equal to the eigenvalues of

$$\frac{1}{n}\left(I_N-nS_{dd}^{-1/2}\bar{d}\bar{d}'S_{dd}^{-1/2}\right)\left(I_N+\frac{1}{n}S_{dd}^{1/2}D'XW^{-1}X'DS_{dd}^{-1/2}\right)$$

which in turn satisfies that $\frac{c_1}{n}\le\lambda_\ell\le\frac{c_2}{n}$ for $\ell=1,\ldots,N-1$ and $c_2\ge c_1>0$ not depending on n. $w_i'w_i=O(P_{ii})$ so Proposition 1 applies when N is fixed and $\min_g T_g\to\infty$. Finally,

$$\begin{aligned}\max_i \mathbb{V}[\hat{\theta}]^{-1}(\tilde{x}_i'\beta)^2 &= O\left(\frac{\max_{g,t}\alpha_g^2+\|x_{gt}\|^2\frac{1}{n}\sum_{i=1}^{n}\|x_{g(i)t(i)}\|^2\sigma_\alpha^2}{N}\right)\\ \max_i \mathbb{V}[\hat{\theta}]^{-1}(\tilde{x}_i'\beta)^2 &= O\left(\frac{\max_{i,j}(x_j'\beta)^2\left(\sum_{\ell=1}^{n}|M_{i\ell}|\right)^2}{N}\right)\end{aligned}$$

and $\sum_{\ell=1}^{n}|M_{i\ell}|=O(1)$ so Proposition 2 applies when $N\to\infty$.

We finish this example with a setup where an unbalanced panel leads to a bias and inconsistency

in $\hat{\theta}_{\text{HO}}$. Consider

$$y_{gt} = \alpha_g + \varepsilon_{gt} \qquad (g = 1, \ldots, N,\ t = 1, \ldots, T_g)$$

where N is even, $(T_g = 2, \mathbb{E}[\varepsilon_{gt}^2] = 2\sigma^2)$ for $g \leq N/2$ and $(T_g = 3, \mathbb{E}[\varepsilon_{gt}^2] = \sigma^2)$ for $g > N/2$, and the estimand is,

$$\theta = \frac{1}{n}\sum_{g=1}^{N} T_g \alpha_g^2 \qquad \text{where } n = \sum_{g=1}^{N} T_g = \frac{5N}{2}.$$

Here we have that $\tilde{A} = I_N/n$ and $\text{trace}(\tilde{A}^2) = N/n^2 = o(1)$ as $n \to \infty$ so the leave-out estimator is consistent. Furthermore,

$$nB_{ii} = P_{ii} = \begin{cases} \frac{1}{2}, & \text{if } i \leq N, \\ \frac{1}{3}, & \text{otherwise,} \end{cases} \qquad \sigma_i^2 = \begin{cases} 2\sigma^2, & \text{if } i \leq N, \\ \sigma^2, & \text{otherwise,} \end{cases}$$

so

$$\mathbb{E}[\tilde{\theta}] - \theta = \sum_{i=1}^{n} B_{ii}\sigma_i^2 = \frac{\sigma^2}{n}\left(N + \frac{N}{2}\right) = \frac{3\sigma^2}{5},$$
$$\mathbb{E}[\hat{\theta}_{\text{HO}}] - \theta = \sigma_{nB_{ii},\sigma_i^2} + S_B \frac{n}{n-N}\sigma_{P_{ii},\sigma_i^2} = \frac{2\sigma^2}{50} + \frac{2}{3} \times \frac{2\sigma^2}{50} = \frac{\sigma^2}{15}.$$

Example 3. $\tilde{A}$ is diagonal with N diagonal entries of $\frac{1}{n}\frac{T_g}{S_{zz,g}}$, so $\lambda_g = \frac{1}{n}\frac{T_g}{S_{zz,g}}$ for $g = 1, \ldots, N$. $\text{trace}(\tilde{A}^2) \leq \frac{\lambda_1}{\min_g S_{zz,g}} \frac{1}{n}\sum_{g=1}^{N} T_g = O(\lambda_1)$. $\max_i w_i' w_i = \max_{g,t} \frac{(z_{gt} - \bar{z}_g)^2}{S_{zz,g}} = o(1)$ when $\min_g S_{zz,g} \to \infty$. Furthermore, $\mathbb{V}[\hat{\theta}]^{-1} = O(\frac{n^2}{N})$, so

$$\mathbb{V}[\hat{\theta}]^{-1} \max_i (\tilde{x}_i'\beta)^2 = O\left(\max_{g,t} \frac{z_{gt}^2 \delta_g^2}{N S_{zz,g}}\right) = o(1),$$

and $M_{i\ell} = 0$ if $g(i) \neq g(\ell)$ so

$$\mathbb{V}[\hat{\theta}]^{-1} \max_i (\tilde{x}_i'\beta)^2 = O\left(\max_g \left(\frac{n\sum_{i:g(i)=g} B_{ii}}{\sqrt{N}}\right)^2\right) = O\left(\max_g \left(\frac{T_g}{\sqrt{N} S_{xx,g}}\right)^2\right) = o(1)$$

both under the condition that $N \to \infty$ and $\frac{\sqrt{N}S_{xx,1}}{T_1}$. Used above:

$$P_{i\ell} = T_{g(i)}^{-1}\mathbf{1}_{\{g(i)=g(\ell)\}} + \frac{(z_{g(i)t(i)} - \bar{z}_{g(i)})(z_{g(i)t(\ell)} - \bar{z}_{g(i)})}{S_{zz,g(i)}}\mathbf{1}_{\{g(i)=g(\ell)\}}$$
$$B_{ii} = \frac{1}{n}\frac{z_{g(i)t(i)} - \bar{z}_{g(i)}}{S_{zz,g(i)}}\frac{T_{g(i)}}{S_{zz,g(i)}}.$$

Finally,

$$\max_i \mathsf{w}_{iq}'\mathsf{w}_{iq} = \max_t \frac{(z_{1t} - \bar{z}_1)^2}{S_{zz,1}} = o(1)$$
$$\mathbb{V}[\hat{\theta}_q]^{-1}\max_i(\tilde{x}_{iq}'\beta)^2 = O\left(\max_{g\geq 2,t}\frac{z_{gt}^2\delta_g^2}{NS_{zz,g}}\right) = o(1),$$
$$\mathbb{V}[\hat{\theta}_q]^{-1}\max_i(\tilde{x}_{iq}'\beta)^2 = O\left(\max_{g\geq 2}\left(\frac{T_g}{\sqrt{N}S_{xx,g}}\right)^2\right) = o(1)$$

under the conditions that $\frac{\sqrt{N}}{T_2}S_{zz,2} \to \infty$ and $S_{zz,1} \to \infty$. Thus, Theorem 1 applies when $\frac{\sqrt{N}}{T_1}S_{zz,1} = O(1)$.

Example 4. Let $\dot{f}_i = (\mathbf{1}_{\{j(g,t)=0\}}, f_i')' = (\mathbf{1}_{\{j(g,t)=0\}}, \mathbf{1}_{\{j(g,t)=1\}}, \ldots, \mathbf{1}_{\{j(g,t)=J\}})'$ and define the following partial design matrices with and without dropping ψ_0 from the model:

$$S_{ff} = \sum_{i=1}^n f_i f_i', \qquad S_{\dot{f}\dot{f}} = \sum_{i=1}^n \dot{f}_i \dot{f}_i', \qquad S_{\Delta f\Delta f} = \sum_{g=1}^N \Delta f_g \Delta f_g', \qquad S_{\Delta\dot{f}\Delta\dot{f}} = \sum_{g=1}^N \Delta\dot{f}_g\Delta\dot{f}_g',$$

where $\Delta\dot{f}_g = \dot{f}_{i(g,2)} - \dot{f}_{i(g,1)}$. Letting $\dot{D}$ be a diagonal matrix that holds the diagonal of $S_{\Delta\dot{f}\Delta\dot{f}}$ we have that

$$E = \dot{D}S_{\dot{f}\dot{f}}^{-1} \quad \text{and} \quad \mathcal{L} = \dot{D}^{-1/2}S_{\Delta\dot{f}\Delta\dot{f}}\dot{D}^{-1/2}.$$

$S_{\Delta\dot{f}\Delta\dot{f}}$ is rank deficient with $S_{\Delta\dot{f}\Delta\dot{f}}\mathbf{1}_{J+1} = 0$ from which it follows that the non-zero eigenvalues of $E^{1/2}\mathcal{L}E^{1/2}$ (which are the non-zero eigenvalues of $S_{\dot{f}\dot{f}}^{-1}S_{\Delta\dot{f}\Delta\dot{f}}$) are also the eigenvalues of $S_{\Delta f\Delta f}(S_{ff}^{-1} + \frac{\mathbf{1}_J\mathbf{1}_J'}{S_{\dot{f}\dot{f},11}})$. Finally, from the Woodbury formula we have that A_{ff} is invertible with

$$A_{ff}^{-1} = n(S_{ff} - n\bar{f}\bar{f}')^{-1} = n\left(S_{ff}^{-1} + n\frac{S_{ff}^{-1}\bar{f}\bar{f}'S_{ff}^{-1}}{1 - n\bar{f}'S_{ff}^{-1}\bar{f}}\right) = n\left(S_{ff}^{-1} + \frac{\mathbf{1}_J\mathbf{1}_J'}{S_{\dot{f}\dot{f},11}}\right),$$

so

$$\lambda_\ell = \lambda_\ell(A_{ff}S^{-1}_{\Delta f\Delta f}) = \frac{1}{\lambda_{J+1-\ell}(S_{\Delta f\Delta f}A^{-1}_{ff})} = \frac{1}{n\lambda_{J+1-\ell}(E^{1/2}\mathcal{L}E^{1/2})}.$$

With E_{jj} constant across j, we have that

$$\frac{\lambda_1^2}{\sum_{\ell=1}^J \lambda_\ell^2} = \frac{\dot\lambda_J^{-2}}{\sum_{\ell=1}^J \dot\lambda_\ell^{-2}} \leq \frac{4}{(\sqrt{J}\dot\lambda_J)^2}$$

since $\dot\lambda_\ell \leq 2$ (Chung, 1997, Lemma 1.7). An algebraic definition of $\mathcal{C}$ is

$$\mathcal{C} = \min_{X\subseteq\{0,\ldots,J\}:\sum_{j\in X}\dot D_{jj}\leq\frac{1}{2}\sum_{j=0}^J \dot D_{jj}} \frac{-\sum_{j\in X}\sum_{k\notin X} S_{\Delta\dot f\Delta\dot f,jk}}{\sum_{j\in X}\dot D_{jj}}$$

and it follows from the Cheeger inequality $\dot\lambda_J \geq 1-\sqrt{1-\mathcal{C}^2}$ (Chung, 1997, Theorem 2.3) that $\sqrt{J}\dot\lambda_J \to \infty$ if $\sqrt{J}\mathcal{C} \to \infty$.