Repairing Locally Misspecified GMM: An Empirical Bayes Approach

Patrick Kline*

UC Berkeley

March 26, 2026

Abstract

Econometric models offer parsimonious but inexact approximations to data-generating processes. This paper studies the generalized method of moments (GMM) when exchangeable specification errors of order $n^{-1/2}$ contaminate the moment conditions. I develop estimators for the mean and variance of these specification errors, establishing their consistency in heavily overidentified models. These hyperparameter estimates are used to develop a feasible first-order bias-corrected estimator of target parameters. I also propose a shrinkage estimator that subtracts a best linear predictor of the first-order estimation error from GMM. Under correct specification, this estimator is asymptotically equivalent to efficient GMM. I derive asymptotic distributions for both estimators and provide variance estimators that enable misspecification-aware frequentist inference. Simulation exercises indicate the procedures can meaningfully improve on standard two-stage least squares estimation when exclusion violations are present. Finally, I revisit the influential study of Angrist and Krueger [1991], considering an instrument set where exchangeability is plausible. Repairing the two-stage least squares estimates of the returns to schooling moves them in the direction of ordinary least squares and reduces sensitivity to the specification of controls.

*This manuscript was prepared for the 2026 Sargan lecture. I am grateful to Isaiah Andrews for helpful feedback on an earlier version of this draft. Claude Opus 4.6 provided outstanding research assistance on this project.

The generalized method of moments (GMM) is a widely used technique for fitting parsimonious semi-parametric models to data. The procedure consists of specifying a function $g : \mathbb{R}^p \to \mathbb{R}^m$ whose zeros define m moment restrictions on a parameter vector $\theta \in \mathbb{R}^p$. A distinguishing feature of GMM, relative to classical method of moments estimators, is that the number of moment restrictions can exceed the number of parameters ($m \geq p$). The case where $m > p$ is known as an overidentified model [Koopmans and Reiersol, 1950]. In practice, GMM models often contain many overidentifying restrictions. For instance, Angrist and Krueger [1991] consider a specification where $m - p = 179$. The primary specification in DellaVigna et al. [2012] has $m - p = 55$, while Gourinchas and Parker [2002] consider a specification with $m - p = 36$.

Foundational work by Sargan [1958] and Hansen [1982] developed a framework for testing these overidentifying restrictions by comparing the J-statistic to its limiting distribution under proper specification. In practice, overidentifying restrictions are typically rejected. Such rejections are often met with a degree of ambivalence, as econometric models are ultimately approximations.[1] However, the nature of these approximation errors is rarely quantified in a precise way. Moreover, model rejections can be difficult to explain to referees, which may explain why J-statistics are rarely reported in modern empirical work [Andrews et al., 2025].

This paper explores how to use the information in overidentifying restrictions to build improved estimates of the model parameters θ that account for specification error. Following Andrews et al. [2017], I work within a local misspecification framework in which the m moment conditions are perturbed by errors that shrink with the sample size. Unlike past work, I treat these errors as exchangeable random variables. I propose estimators of the mean and variance of the specification errors, the latter of which captures overdispersion in the J-statistic. I then establish consistency of these hyperparameter estimates in an asymptotic regime where both m and the sample size n grow large but $m^2/n \to 0$, thereby ruling out settings where many-instruments biases become first order [Altonji and Segal, 1996, Newey and Smith, 2004]. This moderately overidentified regime arguably captures a large share of settings where GMM is employed.

The proposed hyperparameter estimates are used to develop two feasible "misspecification-aware" estimators of structural parameters. First, a plug-in bias correction that subtracts an estimate of the GMM estimator's first-order misspecification bias. Second, a shrinkage adjustment that subtracts a best linear predictor of the first-order estimation error from the GMM estimator. Under correct specification, the shrinkage estimator is asymptotically equivalent to efficient GMM. With misspecification, the shrinkage adjustment exploits the predictable structure of the specification errors to weakly reduce leading-term estimation risk relative to the plug-in bias correction, a mechanism closely related to the finding of Brown [1990] that im-

[1] This ambivalence is not unique to the J-test. Thomas Sargent recalls that "Our good friend Robert E. Lucas, Jr. told us in the early 1980s that our likelihood ratio tests and moment matching tests were rejecting too many good models." [Hansen and Sargent, 2014]

proved estimation of a high-dimensional nuisance component can lower risk for a low-dimensional target. I derive asymptotic distributions for both estimators and provide variance estimators for misspecification-aware frequentist inference.

Two recent working papers take related approaches. One is Andrews et al. [2024] who propose an approach to inference on structural parameters that is valid when priors over moment misspecification are rotation invariant. The distributional assumptions on specification errors considered here strictly nest those of Andrews et al. [2024] because rotation invariant distributions are exchangeable but the converse is not true. In particular, rotation invariant priors must have mean zero, a restriction I relax. A second related paper is Chernozhukov et al. [2025] who propose a quasi-Bayes approach to GMM in which priors are specified over both the structural parameters and the specification errors. While their Bayesian calibration and coverage statements are contingent on an assumed prior over misspecification (e.g., Gaussianity), the present paper only relies on exchangeability of the specification errors for consistent estimation of the leading bias and variance terms.

Though the framework developed here offers new opportunities for misspecification-aware estimation and inference, it also imposes greater demands on empirical researchers. Moment conditions need to be scaled carefully to ensure violations of the nominal model are plausibly exchangeable. These scaling decisions are inherently context dependent, requiring commitment to a particular model of misspecification. Ironically, the exchangeability framework is itself likely to be misspecified in most empirical applications. Nonetheless, the framework plausibly offers a close enough approximation to many research designs for the methods proposed here to be worth reporting as a supplement to standard GMM diagnostics such as the J-statistic.

I illustrate the methods in a series of simulation exercises, finding that both corrections can yield sizable risk improvements even when the degree of overidentification is modest. To assess the theory's asymptotic predictions, I simulate a setting where the number of moments grows with the sample size. The simulations confirm that tests based on the proposed misspecification-aware variance estimators have correct asymptotic size.

Finally, I revisit the influential study of Angrist and Krueger [1991], examining schooling instruments based on season of birth by state of birth interactions that plausibly satisfy exchangeability. I find that moment conditions based on these instruments exhibit substantial mean bias. The changing direction of the bias across control sets is consistent with exogeneity failures of the sort highlighted by Rosenzweig and Wolpin [2000] and Buckles and Hungerman [2013]. The corrections consistently shift the two-stage least squares (TSLS) estimates in the direction of ordinary least squares (OLS), substantially narrowing the dispersion of point estimates across alternative sets of controls. The corrected estimates also exhibit larger standard errors than TSLS, reflecting the misspecification-aware nature of the proposed variance estimators.

Section 1 reviews the basics of GMM under proper specification. Section 2 introduces the local misspecification framework and the presumed stochastic properties of the moment specification errors. Section 3 studies identification of mean and variance of the specification errors. Section 4 proposes estimators of these hyperparameters of the specification error distribution and establishes their consistency. Section 5 discusses estimators of target parameters that account for the specification errors in different ways. Section 6 develops distribution theory for these estimators enabling misspecification-robust inference. Section 7 evaluates the finite sample performance of the estimators in a series of simulation exercises. Section 8 revisits the study of Angrist and Krueger [1991]. Section 9 concludes.

1 GMM preliminaries

Suppose there exists a unique parameter vector θ_0 obeying $g(\theta_0) = 0$, where the population moment condition $g : \mathbb{R}^p \to \mathbb{R}^m$ is continuously differentiable. Let the $m \times p$ matrix G denote the Jacobian $\nabla_\theta g(\theta)$ evaluated at θ_0. Identification of θ_0 requires that G has rank p.

GMM estimates are obtained by constructing empirical moment conditions $\hat{g}(\theta)$ that converge in probability to $g(\theta)$. The GMM estimator $\hat{\theta}$ minimizes the quadratic form $\hat{g}(\theta)' \hat{W} \hat{g}(\theta)$ where $\hat{W}$ is some positive definite weighting matrix that converges in probability to a nonsingular $m \times m$ matrix W. Letting $\hat{G}$ denote the Jacobian of $\hat{g}(\theta)$ evaluated at $\hat{\theta}$, the GMM estimator satisfies the first order condition $\hat{G}'\hat{W}\hat{g}\left(\hat{\theta}\right) = 0$.

It is typically assumed that the scaled empirical moments, when evaluated at the true parameter vector θ_0, converge to normals:

$$\sqrt{n}\hat{g}(\theta_0) \xrightarrow{d} N(0, V).$$

Standard arguments [Newey and McFadden, 1994] yield the influence function representation of GMM estimators

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\left(G'WG\right)^{-1} G'W\sqrt{n}\hat{g}(\theta_0) + o_p(1).$$

Hence, the GMM estimator is, to first order, a linear combination of normals. Andrews et al. [2017] term these combination weights the "sensitivity matrix"

$$\Lambda := -\left(G'WG\right)^{-1} G'W,$$

which summarizes the local influence of the empirical moment conditions on the parameter estimates.

Note that premultiplying a vector of moments by $-\Lambda$ amounts to a weighted least squares regression using the columns of the Jacobian G as regressors. It will be useful to consider the corresponding "hat" matrix

$H := -G\Lambda$. Premultiplying a vector of moments by H gives predicted values of the moment conditions generated from a weighted least squares fit. The sample analogues of Λ and H will be denoted by $\hat{\Lambda} = -\left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}$ and $\hat{H} = -\hat{G}\hat{\Lambda}$, respectively.

2 A model of local misspecification

To introduce the possibility of model misspecification, I follow Andrews et al. [2017] in allowing the distribution of the moment conditions to be contaminated by a vector of specification errors $b_n/\sqrt{n}$ that shrinks with the sample size. Define

$$r_n := \sqrt{n}\hat{g}(\theta_0) = b_n + \varepsilon_n.$$

The vector b_n captures specification errors, while ε_n captures noise in the moment conditions attributable to sampling variation. For any finite sample size n, the specification errors can generate bias, but the central idea of the local misspecification framework is that $b_n/\sqrt{n} = o_p(1)$, which ensures that GMM remains consistent for the true parameter vector of interest θ_0.

While Andrews et al. [2017] treat the vector $b_n \in \mathbb{R}^m$ as fixed, I will view it as random with exchangeable coordinates. Formally, exchangeability means that the distribution of b_n is invariant to any permutation of its m entries. This is a reasonable perspective to take when i) the moment conditions all take the same basic form — i.e., they are all measured in the same units and involve averages over similar mathematical objects — and ii) one does not know which moment conditions are likely to be misspecified a priori. I will have more to say about i) below but ii) is clearly a situation with which empirical researchers often find themselves confronted. The following assumption provides an asymptotic version of this perspective.

Assumption 1. *The random vectors (b_n, ε_n) obey the following conditions:*

i) $(b_n, \varepsilon_n) \xrightarrow{d} (b, \varepsilon)$ with $b \perp \varepsilon$ and $\varepsilon \sim N(0, V)$,

ii) The components $\{b_j\}_{j=1}^m$ of b are exchangeable with $E[b_j] = \bar{\mu}$, $Var[b] = 1_m 1_m' \tau^2 + I_m \bar{\sigma}^2$, and $E\left[b_j^4\right] < \infty$

Assumption 1.i imposes that specification errors are asymptotically independent of the noise in the moment conditions. This is a substantive economic restriction that could be violated if, for example, a data-dependent moment selection rule induces correlation between b and ε. Assumption 1.ii restricts the entries of b to be exchangeable with bounded fourth moment. This too is a strong assumption, which I discuss in greater detail below. It would be violated if, for example, less precise moment conditions are more likely to be misspecified, so that the distribution of b_j depends on the moment variance V_{jj}. When $\tau^2 > 0$, the specification errors are correlated.

Without loss of generality, one can represent b as a location-scale mixture of exchangeable random variables

$$\begin{aligned} b &= (\bar{\mu} + \upsilon) 1_m + \varsigma \bar{\sigma} \xi \\ &\equiv \mu 1_m + \sigma \xi. \end{aligned}$$

Here, $\upsilon \in \mathbb{R}$ is a "common shock" to the location of the moment conditions with mean zero and variance τ^2. The standardized specification errors $\xi \in \mathbb{R}^m$ obey $Var[\xi] = E[\xi\xi'] = I_m$ but may not be mutually independent. The quantity $\varsigma \in \mathbb{R}$ is a common shock to the scale of these errors obeying $E\left[\varsigma^2\right] = 1$.

One way to motivate this representation is to consider a researcher who randomly samples moment conditions from a broader super-population of misspecified moment conditions. The random variable $\mu = \bar{\mu} + \upsilon$ is the conditional mean of the specification errors among the sampled moments given the draw υ, while the constant $\bar{\mu}$ captures the average error that would be present in a broader super-population of moment conditions that might have been considered. Likewise, the parameter $\bar{\sigma}$ gives the scale of specification errors in the super-population, while $\sigma = \varsigma\bar{\sigma}$ gives their conditional scale among the sampled moments given the draw ς.

With a single set of moment conditions, it won't be possible to separately identify $\bar{\mu}$ from υ or ς from $\bar{\sigma}$. Fortunately, it is not essential to do so. Since the goal of this analysis will be to improve GMM estimates based on the moment conditions actually chosen by the researcher, it is natural to be concerned with the mean and variance of the conditional asymptotic bias of GMM

$$E\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right) \mid \upsilon, \varsigma\right] = \mu\Lambda 1_m + o(1), \quad Var\left[E\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right) \mid b\right] \mid \upsilon, \varsigma\right] = \sigma^2\Lambda\Lambda' + o(1).$$

Accordingly, I will seek estimates of μ and σ^2 and use them to construct improved estimators that take into account the conditional influence of misspecification errors. With this motivation in mind, I subsequently condition on (υ, ς), which implies μ and σ are fixed but leaves the entries of b potentially dependent. To avoid cumbersome notation, the conditioning on (υ, ς) will be left implicit wherever possible in the text.

2.1 Examples

It is useful to consider some examples in order to illustrate both how Assumption 1 can be violated and how to design moment conditions where it can plausibly be satisfied. I start with an example where $p = 1$.

Example 1 (Equal means)**.** Let Y_n be an $m \times 1$ vector of independent sample means with $Var(Y_n) = diag(v_1, \ldots, v_m)/n = V/n$, where n is the total sample size. To map to the local misspecification framework,

I index the corresponding m-vector of population means ϑ_n by the sample size, the jth entry of which I denote ϑ_{nj}.

The researcher considers a model that restricts all m population means to equal a common scalar θ. The resulting m-vector of population moment conditions takes the form $g(\theta) = \vartheta_n - \theta 1_m$, which implies the population Jacobian is $G = -1_m$. To think about local deviations from this model, suppose that there exists a true θ_0 and that the deviations $\vartheta_{nj} - \theta_0$ are random disturbances of order $O_p\left(n^{-1/2}\right)$. Thus, GMM will be consistent and its scaled population moment conditions equal the specification errors $b = \sqrt{n}(\vartheta_n - \theta_0 1_m)$, with entries b_j.

If the $\{b_j\}_{j=1}^m$ are exchangeable with mean $\bar{\mu}$, then Assumption 1 is satisfied. In contrast, Assumption 1 will be violated if either $E[b_j] = f_j(V, W)$ or $Var(b_j) = f_j(V, W)$ for some function f_j that varies with j. For example, if noisier means have larger deviations then one might expect $E[b_j] \propto \sqrt{v_j}$, entailing a simple version of the precision dependence problem studied by Chen [2022].

The following example considers a model with $p = 2$.

Example 2 (Two groups of means)**.** Returning to the case with m means, suppose $m \geq 4$ is even and let $B = I_2 \otimes 1_{m/2}$ be a matrix of indicators dividing the means into two "blocks" of equal size. The researcher's model imposes that the means in each block are equal. This yields moment conditions $g(\theta) = \vartheta_n - B\theta$, where now $\theta \in \mathbb{R}^2$ and $G = -I_2 \otimes 1_{m/2}$.

Define $b = \sqrt{n}(\vartheta_n - B\theta_0)$. As before, if the $\{b_j\}_{j=1}^m$ are exchangeable with mean $\bar{\mu}$, then Assumption 1 is satisfied. In contrast, Assumption 1 will be violated if the distribution of b depends on G. For example, suppose the first block of specification errors $\{b_j\}_{j=1}^{m/2}$ are exchangeable with mean $\bar{\mu}_1$ and variance $\bar{\sigma}_1^2$, the second block $\{b_j\}_{j=m/2+1}^m$ are exchangeable with mean $\bar{\mu}_2$ and variance $\bar{\sigma}_2^2$, and either $\bar{\mu}_1 \neq \bar{\mu}_2$ or $\bar{\sigma}_1^2 \neq \bar{\sigma}_2^2$. Such violations might be natural if the two blocks correspond to different waves of a survey; see Krueger et al. [2017] for a discussion of such concerns in the context of measuring unemployment rates.

As the above example illustrates, the plausibility of the exchangeability assumptions depends critically on the structure of the Jacobian G. The following example considers an overidentified instrumental variables model in which G takes a more complex form.

Example 3 (Overidentified IV)**.** A researcher has cross-sectional data on $m > 2$ instruments $Z_{i1}, \ldots, Z_{im}$ for a binary endogenous variable D_i, where i indexes individual level observations. Collect the instruments into the vector $Z_i := (Z_{i1}, \ldots, Z_{im})'$ and denote the outcome by Y_i. The triples $\{Y_i, D_i, Z_i'\}_{i=1}^n$ are presumed to be *i.i.d.* samples from a data generating process (DGP) $\mathbf{P}_n$ indexed by the sample size to capture local misspecification. I use $E_n[\cdot]$ to denote expectations with respect to this drifting DGP.

The researcher's model stipulates the population moment restrictions $g(\theta) = E_n[(Y_i - \theta D_i) Z_i]$ for scalar θ, implying $G = -E_n[D_i Z_i]$. Now define $\vartheta_n := E_n[Y_i Z_i]$ and suppose $\vartheta_n = \theta_0 E_n[D_i Z_i] + b/\sqrt{n}$. Thus, we can equivalently write the misspecified moment restrictions $g(\theta) = \vartheta_n - \theta E_n[D_i Z_i]$, which yields $g(\theta_0) = b/\sqrt{n}$.

As before, if the $\{b_j\}_{j=1}^m$ are exchangeable with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$, then Assumption 1 is satisfied. However, if either of the first two moments of b_j are functions of G, W, or V, then Assumption 1 will be violated. For instance, suppose the TSLS weight matrix $W = E[Z_i Z_i']^{-1}$ is employed and that $b_j = \sum_k E_n[Z_{ij} Z_{ik}] \gamma_k$, where γ_k is the direct effect of instrument k on the outcome. The cross-moment matrix $E_n[Z_i Z_i']$ makes the distribution of b_j depend on W. If, in addition, stronger instruments tend to have larger $|\gamma_j|$, then $E[b_j]$ and $\text{Var}[b_j]$ also depend on G, violating Assumption 1.

Example 4 (Overidentified IV (cont.))**.** The above example assumed that $E_n[(Y_i - \theta_0 D_i) Z_i] = b/\sqrt{n}$, where the elements of b are exchangeable. Another plausible model of exclusion violations posits that, for some vector $\gamma \in \mathbb{R}^m$ of direct effects

$$\vartheta_n = \theta_0 E_n[D_i Z_i] + \frac{1}{\sqrt{n}} E_n[Z_i Z_i'] \gamma \Rightarrow E_n\left[\left(Y_i - \theta_0 D_i - \frac{1}{\sqrt{n}} Z_i'\gamma\right) Z_i\right] = 0,$$

where the direct effects $\{\gamma_j\}_{j=1}^m$ are exchangeable and independent of Z_i. In this case, $b = E_n[Z_i Z_i'] \gamma$ is generally not exchangeable unless $E_n[Z_i Z_i'] \propto I_m$. This difficulty can be circumvented by rescaling the posited moment conditions as

$$g(\theta) = E_n[Z_i Z_i']^{-1} E_n[(Y_i - \theta D_i) Z_i] = 0,$$

which imply that $g(\theta_0) = \gamma/\sqrt{n}$ is exchangeable.

As the contrast between these last two examples illustrates, the researcher must think carefully about what aspect of the model is misspecified. While in Example 3 the exchangeable errors were presumed to be additive shifts in an orthogonality condition, Example 4 provides a plausible mechanism by which the errors are proportional to instrument scale.

Remark 1. It is tempting to address these scaling issues by reparametrizing the moment conditions relative to the weighting matrix – e.g., by modeling the specification errors as $W^{1/2}\sqrt{n} g(\theta_0) = b$. While such an approach could be pursued, doing so would risk confusing questions of efficiency with scaling. For instance, under the form of excludability violations discussed in Example 4, the proper scaling amounts to GMM estimation with a weighting matrix of $E_n[Z_i Z_i']^{-2}$, which is the square of the usual TSLS weighting. In what follows, I will rely on the researcher to scale the moment conditions directly in a form obeying $g(\theta_0) = b/\sqrt{n}$

in order to separate the choice of weighting matrix from other considerations.

2.2 Implications of Assumption 1

Let $S := \sigma^2 I_m$ and recall that all statements are implicitly conditional on υ and ς. The restrictions in Assumption 1 imply

$$r_n \xrightarrow{d} b + \varepsilon \equiv r, \quad E[r] = \mu 1_m, \quad Var[r] = V + S \equiv \Sigma.$$

Hence, one can write

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \Lambda r + o_p(1).$$

It follows that the bias in GMM conditional on b is

$$E\left[\hat{\theta} \mid b\right] - \theta_0 = \frac{1}{\sqrt{n}}\Lambda b + o\left(n^{-1/2}\right).$$

Averaging over b, the marginal bias is

$$E\left[\hat{\theta} - \theta_0\right] = \frac{\mu}{\sqrt{n}} \cdot \Lambda 1_m + o\left(n^{-1/2}\right).$$

Thus, the leading term of the bias in $\hat{\theta}$ is $\frac{\mu}{\sqrt{n}} \cdot \Lambda 1_m$.

The marginal variance of the GMM estimator splits into two terms

$$Var\left[\hat{\theta}\right] = \underbrace{\frac{1}{n}\Lambda V \Lambda'}_{\text{moment uncertainty}} + \underbrace{\frac{1}{n}\Lambda S \Lambda'}_{\text{specification error}} + o\left(n^{-1}\right).$$

In Section 4, I will propose empirical Bayes estimators that mitigate both the bias and the variance attributable to specification errors. These estimators leverage estimates of the moments of those errors, which I will view as hyperparameters in what follows.

3 Identification of hyperparameters

This section studies identification of $\left(\mu, \sigma^2\right)$. Recall that the GMM estimator sets p linear combinations of the sample moments equal to zero. Define the sample annihilator matrix $\hat{M} := I_m - \hat{H}$ and its population analogue

$$M := I_m - H = I_m + G\Lambda.$$

Assumption 1 implies the scaled moment conditions can be written

$$\hat{r} := \sqrt{n}\hat{g}\left(\hat{\theta}\right) = \hat{M}\left(b + \varepsilon\right) + o_p\left(1\right) = Mr + o_p\left(1\right).$$

Intuitively, the process of fitting GMM masks p linear combinations of the b_j's that lie in the nullspace of M. The exchangeability assumption allows us to infer the distribution of these hidden linear combinations from the $m - p$ identified linear combinations that lie in M's column space.

An oracle that knows the hyperparameters $\left(\mu, \sigma^2\right)$ and the population residual Mr can reduce the first-order estimation error in $\hat{\theta}$ by subtracting the best linear predictor of r given Mr

$$L\left[r \mid Mr; \mu, \sigma^2\right] = \mu 1_m + \Sigma M'\left(M\Sigma M'\right)^{+}\left(Mr - \mu M 1_m\right).$$

I revisit this thought experiment in Section 5.2, which proposes a shrinkage estimator of θ_0.

3.1 Moment equations

Assumption 1 implies

$$E\left[\hat{r}\right] = \mu M 1_m + o\left(1\right), \quad Var\left[\hat{r}\right] = M\Sigma M' + o\left(1\right).$$

Provided $\left(M1_m\right)'\left(M1_m\right) > 0$, it follows that

$$E\left[\frac{\left(M1_m\right)'\hat{r}}{\left\|M1_m\right\|^2}\right] = \mu + o\left(1\right), \quad E\left[\frac{\left\|\hat{r} - \mu M 1_m\right\|^2 - \operatorname{tr}\left(MVM'\right)}{\operatorname{tr}\left(MM'\right)}\right] = \sigma^2 + o\left(1\right). \tag{1}$$

Note that $\frac{\left(M1_m\right)'\hat{r}}{\left\|M1_m\right\|^2}$ is the coefficient from an OLS regression of $\hat{r}$ on $M1_m$. The ratio $\frac{\left\|\hat{r} - \mu M 1_m\right\|^2 - \operatorname{tr}\left(MVM'\right)}{\operatorname{tr}\left(MM'\right)}$ is a degrees-of-freedom adjusted estimator of the variance of the specification errors, obtained by subtracting the expected sampling noise from the mean squared residual after demeaning by the population μ. These equations motivate the hyperparameter estimators that will be proposed in the next section.

3.2 Identification failures

In overidentified models, $\operatorname{tr}\left(MM'\right) > 0$. Hence, σ^2 is generically identified given $m > p$. In contrast, some overidentified models imply $M1_m = 0$, leading to non-identification of μ. This problem arises whenever 1_m lies in the column space of G.

3.2.1 Location models

Recall that in Example 1 $G = -1_m$, implying the mean bias loads directly onto the parameter of interest: $E\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right)\right] = \mu + o(1)$. In Example 2 $G1_2 = -1_m$ and the mean specification error shifts both parameters equally: $E\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right)\right] = \mu 1_2 + o(1)$. In both cases, the hyperparameter μ is absorbed by the GMM fitting process because it is collinear with the parameter being estimated.

3.2.2 Intercepts in linear models

In overidentified linear models, μ typically remains identified even when an intercept is included. Consider an extension of Example 3 where $\theta = (\alpha, \beta)'$, with α capturing an intercept and β a slope. Moreover, an intercept is added to the m fundamental instruments Z_i so that the moment conditions take the form $g(\theta) = E_n\left[(Y_i - (1, D_i)\theta)(1, Z_i')'\right]$. In this case

$$G = -\begin{bmatrix} 1 & E_n[D_i] \\ E_n[Z_i] & E_n[D_i Z_i] \end{bmatrix}.$$

Unless $(1 - E_n[Z_{ij}])/Cov_n(D_i, Z_{ij})$ is constant across j, the column space of G will not span 1_{m+1}.

In contrast, identification fails in an extension of Example 3 with $\theta = (\alpha, \beta)'$ and mutually exclusive binary instruments $Z_{ij} \in \{0, 1\}$ obeying $\sum_{j=1}^{m} Z_{ij} = 1$, a design that arises often in the recent literature on judge IV [Chyn et al., 2025]. To understand the problem, consider the scaled moment conditions

$$g(\theta) = E_n[Z_i Z_i']^{-1} E_n[(Y_i - (1, D_i)\theta) Z_i].$$

Letting $q = E_n[Z_i]$, the Jacobian takes the form

$$G = -\text{diag}(q)^{-1}(E_n[Z_i], E_n[D_i Z_i]) = -\left(1_m, diag(q)^{-1} E_n[D_i Z_i]\right).$$

From the first column, 1_m lies in the span of G. Therefore $M 1_m = 0$.

3.2.3 Which parameters are biased?

Non-identification of μ implies that some linear combination of the entries in $\hat{\theta}$ exhibits irreducible bias of order $O\left(n^{-1/2}\right)$. However, other parameters may be unbiased. In the extension of Example 3, μ loads directly on the intercept parameter α. However, β is insensitive to μ, yielding no leading-term bias in the

GMM point estimate $\hat{\beta}$:

$$E\left[\sqrt{n}\left(\hat{\beta}-\beta_0\right)\right]=\mu\Lambda_\beta 1_m+o\left(1\right)=o\left(1\right),$$

where Λ_β is the row of Λ corresponding to β. The second equality uses $\Lambda_\beta 1_m=0$, which holds because 1_m lies in the column space of G and $\Lambda G=-I_p$.

Though GMM exhibits no leading-term bias for β in this example, it does exhibit conditional bias $E\left[\sqrt{n}\left(\hat{\beta}-\beta_0\right)\mid b\right]=\Lambda_\beta b+o\left(1\right)$. Averaging over the exchangeable specification errors, this conditional bias contributes variance $\sigma^2\Lambda_\beta\Lambda_\beta'$ to the asymptotic distribution of $\hat{\beta}$:

$$Var\left[\sqrt{n}\left(\hat{\beta}-\beta_0\right)\right]=\underbrace{\Lambda_\beta V\Lambda_\beta'}_{\text{sampling error}}+\underbrace{\sigma^2\Lambda_\beta\Lambda_\beta'}_{\text{specification error}}+o\left(1\right).$$

In this scenario, the job of "repairing" GMM simplifies to reducing the noise in $\hat{\beta}$ attributable to specification error using estimates of σ^2, a task that can be accomplished via empirical Bayes shrinkage.

4 Estimating hyperparameters

This section considers how to estimate the hyperparameters $\left(\mu,\sigma^2\right)$ in an environment where m and n grow jointly, while p remains fixed. Intuitively, these results leverage the high degree of overidentification $m-p$ found in many empirical studies, which provide opportunities to detect overdispersion in the empirical moment conditions attributable to misspecification.

The results in this section are established under a joint asymptotic framework in which both m and n grow to infinity simultaneously, with m growing more slowly than $\sqrt{n}$. This framework allows us to state explicit rate conditions under which the feasible estimators approximate their oracle counterparts. I collect the regularity conditions used throughout below.

Assumption 2. *(Asymptotic Regime and Regularity). The following conditions hold as $n\to\infty$:*

i) $m=m(n)\to\infty$ with $m^2/n\to 0$, and p is fixed.

ii) $\|W\|_2=1$ and $\|V\|_2=1$ (normalizations). The row sums of W are uniformly bounded: $\max_{1\le j\le m}\sum_k|W_{jk}|\le C$.

iii) The first-stage estimators satisfy, uniformly in m: $\|\hat{M}-M\|_2=O_p(\sqrt{m/n})$, $\|\hat{V}-V\|_2=O_p(\sqrt{m/n})$, $\|\hat{\Lambda}-\Lambda\|_2=O_p(1/\sqrt{n})$, and the linearization remainder satisfies $\|\sqrt{n}\hat{g}(\hat{\theta})-\sqrt{n}\hat{g}(\theta_0)-\hat{G}\sqrt{n}(\hat{\theta}-\theta_0)\|=O_p(\sqrt{m/n})$.

iv) The entries of G are uniformly bounded: $|G_{jk}|\le K$ for some $K<\infty$. Additionally, $m^{-1}G'W1_m\to 0$, the eigenvalues of $m^{-1}G'WG$ are bounded away from 0 and ∞, and $m^{-1}1_m'W1_m\to c\in(0,\infty)$.

v) The components of ε are weakly dependent across the moment index j, with uniformly bounded fourth moments and $\sum_k |V_{jk}|$ uniformly bounded.

vi) The nonzero eigenvalues of $M\Sigma M'$ are bounded away from zero uniformly in m.

vii) The exchangeable mean zero specification errors ξ satisfy, for any sequence of symmetric $m \times m$ matrices A_m with $\|A_m\|_2 < C$,

$$E\left[\left(\xi' A_m \xi - tr(A_m)\right)^2\right] \leq C' \cdot tr(A_m^2).$$

Condition (i) specifies that the number of moment conditions grows with the sample size but at a rate slower than $\sqrt{n}$, ensuring that first-stage estimation errors remain asymptotically negligible. When moment conditions are added at a faster rate, a many-instruments bias emerges that not only compromises GMM but the sensitivity-based corrections that will be proposed. Thus, the asymptotic results that follow apply to use cases where models are heavily overidentified, but not so complex that techniques for finite dimensional modeling break down. In Section 8, I explore a variant of a specification considered by Angrist and Krueger [1991] that fits these criteria.

The first part of condition (ii) is without loss of generality: GMM is invariant to rescaling the moment conditions by a scalar c (which rescales V by c^2) and to rescaling W by a scalar d; these two independent scalars allow both normalizations to be imposed simultaneously. Note that $\|WV\|_2 \leq \|W\|_2 \|V\|_2 = 1$ follows immediately. The bounded row sums condition holds automatically when W is diagonal (each row sum equals the diagonal entry, bounded by $\|W\|_2 = 1$) and more generally for any W whose entries do not concentrate on a single row.

Condition (iii) requires that the first-stage estimators converge uniformly in m. The $m \times m$ matrix $\hat{V}$ converges to V at rate $O_p(\sqrt{m/n})$ by standard concentration inequalities for sample covariance matrices. The projection matrix $\hat{M}$ inherits the same rate because its components $\hat{G}$ and $\hat{W}$ each converge at this rate. The sensitivity matrix $\hat{\Lambda}$ converges at the faster parametric rate $O_p(1/\sqrt{n})$. The linearization remainder condition ensures that the scaled residuals $\hat{r} = \sqrt{n}\hat{g}(\hat{\theta})$ are well-approximated by $\hat{M}\sqrt{n}\hat{g}(\theta_0)$ under joint asymptotics; this holds automatically for linear moment conditions and follows from uniformly bounded Hessians in the nonlinear case.

Condition (iv) requires uniformly bounded entries of the Jacobian G, which ensures that individual moment conditions do not exert disproportionate influence on the parameter estimates. This entry-wise bound, together with the convergence of $m^{-1}G'WG$ to a positive definite limit and the vanishing of $m^{-1}G'W1_m$, implies that $(M1_m)_j \to 1$ uniformly — the key to consistent estimation of the mean specification error μ. Combined with the normalization $\|W\|_2 = 1$, it also gives $\|M\|_2 = O(1)$. Condition (v) regulates the

moment noise: uniformly bounded row sums of V control pairwise dependence. Condition (vi) regulates the signal-to-noise ratio as the number of moments grows.

Finally, recall that Assumption 1 restricts the entries of b to be mutually uncorrelated conditional on υ and ς but allows other forms of dependence. Condition (vii) restricts the nature of this remaining dependence in a way that ensures quadratic forms in the centered and scaled errors $\xi = (b - \mu 1_m)/\sigma$ concentrate around their expectations at the rate implied by their trace. It can be shown that this high-level condition is satisfied whenever the law governing ξ is either *i.i.d.* or rotation invariant. Thus, the allowable set of distributions for b strictly nests the family considered by Andrews et al. [2024].

4.1 Estimating μ

Suppose that

$$\left(\hat{M}1_m\right)'\left(\hat{M}1_m\right) > 0 \quad a.s.$$

As discussed in Section 3.2, this condition fails if a linear combination of the columns of $\hat{G}$ equals the unit vector 1_m.

A plug-in estimator of μ motivated by (1) is the regression coefficient

$$\hat{\mu} = \frac{\left(\hat{M}1_m\right)'\hat{r}}{\left\|\hat{M}1_m\right\|^2}.$$

Under Assumption 2, the first-stage estimator $\hat{M}$ converges to its population counterpart, and the estimation error in $\hat{\mu}$ can be controlled jointly in m and n. The following Lemma establishes the convergence rate.

Lemma 1. *Under Assumptions 1 and 2,* $\|M1_m\|^2 \to \infty$ *and*

$$\hat{\mu} - \mu = O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right).$$

Proof. See appendix. □

Thus, $\hat{\mu}$ is consistent for μ when m grows more slowly than n.

Remark 2. The conclusion $\|M1_m\|^2 \to \infty$ reveals that the high-level conditions of Assumption2 characterize an asymptotic regime where $\hat{\mu}$ is *strongly identified*. An interesting question for future work is how to characterize the behavior of $\hat{\mu}$ under weak identification, a setting in which $\|M1_m\|^2 = O(1)$. Though $\hat{\mu}$ is

no longer consistent for μ in this case, it still converges to a well-defined asymptotic distribution that could be relevant for inference in some environments.

4.2 Estimating σ^2

From (1), a plug-in estimator of σ^2 takes the form

$$\widetilde{\sigma^2} = \frac{\left\|\hat{r} - \hat{\mu}\hat{M}1_m\right\|^2 - \operatorname{tr}\left(\hat{M}\hat{V}\hat{M}'\right)}{\operatorname{tr}\left(\hat{M}\hat{M}'\right)},$$

where $\hat{V}$ is a consistent estimator of V. To ensure the estimate is non-negative, one can rely on the truncated estimator

$$\widehat{\sigma^2} = \max\left\{0, \widetilde{\sigma^2}\right\}.$$

4.2.1 Connection to the J-statistic

The estimator $\widetilde{\sigma^2}$ bears a close connection to the usual J-statistic, which can be written

$$\hat{J} = \hat{r}'\hat{V}^{-1}\hat{r},$$

where $\hat{r} = \sqrt{n}\hat{g}\left(\hat{\theta}\right)$ and $\hat{\theta}$ is constructed using $\hat{W} = \hat{V}^{-1}$.

In the special case where $\hat{W} = \hat{V} = I$, one can write

$$\widetilde{\sigma^2} = \frac{\hat{J} - (m-p) - \hat{\mu}^2\, 1_m'\hat{M}'\hat{M}1_m}{m-p}.$$

Note that $m-p$ is the expected value of $\hat{J}$ under correct specification. The additional term $-\hat{\mu}^2\, 1_m'\hat{M}'\hat{M}1_m$ adjusts for the value of $\hat{J}$ expected to arise when $\mu = \hat{\mu}$ and $\sigma^2 = 0$. Hence, in this case, $\widetilde{\sigma^2}$ captures overdispersion in the moment conditions attributable to specification error.

Because $\widetilde{\sigma^2}$ weights moment deviations equally, this intuition does not extend directly to other choices of $\hat{W}$ and $\hat{V}$, as the J-statistic weights moments by their precision. However, if one restricts attention to the optimally weighted case in which $\hat{W} = \hat{V}^{-1}$, the following alternative estimator of σ^2 can also be shown to be consistent:

$$\widetilde{\sigma_J^2} := \frac{\hat{J} - (m-p) - \hat{\mu}^2\, 1_m'\hat{M}'\hat{V}^{-1}\hat{M}1_m}{\operatorname{tr}(\hat{V}^{-1}\hat{M}\hat{M}')}.$$

Note that when $\hat{V} = I$, $\widetilde{\sigma_J^2} = \widetilde{\sigma^2}$.

It can be shown that the minimum variance quadratic-form estimator weights by the unknown matrix

$(M\Sigma M')^+$, where + denotes an MP inverse. Thus, while the connection of $\widetilde{\sigma_J^2}$ to optimally-weighted GMM is appealing, weighting squared moment deviations by $\hat{V}^{-1}$ need not yield uniformly more precise estimates of σ^2 than $\widetilde{\sigma^2}$. Moreover, inverse variance weighting can yield finite sample biases due to correlation between the weights and moment conditions [Altonji and Segal, 1996, Newey and Smith, 2004].

4.2.2 Consistency

Note that

$$tr\left(\hat{M}\hat{M}'\right) = \sum_{i,j=1}^{m} \hat{M}_{ij}^2 \geq \frac{tr\left(\hat{M}\right)^2}{m} = \frac{(m-p)^2}{m},$$

where the inequality follows from Cauchy-Schwarz and the final equality follows from the fact that

$$tr\left(\hat{G}\left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}\right) = tr\left(\left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}\hat{G}\right) = tr\left(I_p\right) = p.$$

Thus, the denominator of $\widetilde{\sigma^2}$ must grow with m.

The following Lemma establishes the convergence rate of the variance estimator under joint asymptotics.

Lemma 2. *Under Assumptions 1 and 2,*

$$\widehat{\sigma^2} - \sigma^2 = O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right).$$

Proof. See appendix. □

It follows immediately from this Lemma that $\hat{S} = \widehat{\sigma^2} I_m$ is consistent for S and $\hat{\Sigma} = \hat{V} + \hat{S}$ is consistent for Σ.

Remark 3. While Assumption 2 stipulates that $m^2/n \to 0$, the rates in Lemmas 1 and 2 require only $m/n \to 0$, under which both the oracle error $O_p(1/\sqrt{m})$ and the first-stage error $O_p(\sqrt{m/n})$ vanish. The stronger condition $m^2/n \to 0$ is driven by the $m \times m$ matrix estimation problems that arise in the feasible bias correction and shrinkage estimators of Section 5, for which the $O(m^2)$ entries of $\hat{M}$ must be estimated precisely enough that the approximation error is $o_p\left(1/\sqrt{mn}\right)$.

5 Bias correction and shrinkage

Recall from before that

$$E\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right)\right] = \mu\Lambda 1_m + o\left(1\right).$$

An oracle that knows μ can construct the bias-corrected estimator

$$\theta^*_{BC} = \hat{\theta} - \frac{\mu}{\sqrt{n}} \cdot \hat{\Lambda} 1_m,$$

which motivates consideration of the feasible bias-corrected estimator

$$\hat{\theta}_{BC} := \hat{\theta} - \frac{\hat{\mu}}{\sqrt{n}} \cdot \hat{\Lambda} 1_m.$$

The following proposition establishes the rate at which $\hat{\theta}_{BC}$ converges to θ_0.

Proposition 1. *Under Assumptions 1 and 2, the feasible bias-corrected estimator satisfies:*

$$\hat{\theta}_{BC} - \theta_0 = O_p\left(\frac{1}{\sqrt{mn}}\right) + O_p\left(\frac{\sqrt{m}}{n}\right).$$

The first term reflects both the oracle estimation error $\theta^*_{BC} - \theta_0$ and the cost of estimating μ, the latter being of smaller order under Assumption 2(iv) since $\Lambda 1_m = o(1)$. The second term arises from the randomness of $\hat{\Lambda}$.

Proof. Under the joint asymptotics of Assumption 2, Lemma 1 gives $\hat{\mu} - \mu = O_p(1/\sqrt{m}) + O_p(\sqrt{m/n})$. Because $\|\hat{\Lambda}\|_2$ scales as $O_p(1/\sqrt{m})$ and $\|1_m\| = \sqrt{m}$, the quantity $\|\hat{\Lambda} 1_m\|$ is bounded by $O_p(1)$. Thus, the feasibility cost of estimating μ is:

$$\frac{1}{\sqrt{n}} (\hat{\mu} - \mu) \hat{\Lambda} 1_m = O_p\left(\frac{1}{\sqrt{mn}}\right).$$

Here the $O_p(\sqrt{m/n})$ component of $\hat{\mu} - \mu$ contributes $O_p(\sqrt{m}/n)$ to the product, which is $o_p(1/\sqrt{mn})$ since $m^2/n \to 0$ implies $m/\sqrt{n} \to 0$.

Thus, from estimating μ, the feasible bias correction introduces an error of:

$$\hat{\theta}_{BC} - \theta^*_{BC} = O_p\left(\frac{1}{\sqrt{mn}}\right).$$

This term vanishes because $mn \to \infty$. The oracle error is $\theta^*_{BC} - \theta_0 = (1/\sqrt{n})\Lambda(\sigma\xi + \varepsilon) + O_p(\sqrt{m}/n)$, where $\|\Lambda(\sigma\xi + \varepsilon)\| = O_p(1/\sqrt{m})$ because $\Lambda(\sigma\xi + \varepsilon)$ is mean-zero with variance $\Lambda\Sigma\Lambda' = O(1/m)$. Combining with the feasibility cost:

$$\hat{\theta}_{BC} - \theta_0 = \left(\hat{\theta}_{BC} - \theta^*_{BC}\right) + (\theta^*_{BC} - \theta_0) = O_p\left(\frac{1}{\sqrt{mn}}\right) + O_p\left(\frac{\sqrt{m}}{n}\right).$$

□

While $\hat{\theta}_{BC}$ is asymptotically unbiased, it is not necessarily the "best" estimator in terms of mean squared error. In what follows, I will seek an estimator of θ_0 that improves on $\hat{\theta}_{BC}$ in asymptotic mean squared error.

5.1 Oracle BLP and shrinkage estimator

The best linear predictor of r given Mr is

$$L\left[r \mid Mr; \mu, \sigma^2\right] = \mu 1_m + \Sigma M' \left(M\Sigma M'\right)^{+} \left(Mr - \mu M 1_m\right).$$

This predictor minimizes $E\|r - (a + BMr)\|^2$ over all affine functions $a + BMr$. Under correct specification ($\mu = \sigma^2 = 0$), $\Sigma = V$ and the predictor reduces to $\Pi_V r$, where $\Pi_V \equiv VM' \left(MVM'\right)^{+} M$ is an oblique projection onto the column space of VM'.

Recall that $\hat{\theta} - \theta_0 = \frac{1}{\sqrt{n}}\Lambda r + o_p\left(n^{-1/2}\right)$. An oracle that knows the hyperparameters $\left(\mu, \sigma^2\right)$ — but not Mr — might therefore consider the following linear shrinkage estimator of θ_0:

$$\theta^* := \hat{\theta} - \frac{1}{\sqrt{n}}\hat{\Lambda} L\left[r \mid \hat{M}\hat{r}; \mu, \sigma^2\right].$$

This oracle estimator subtracts the predicted first-order estimation error from $\hat{\theta}$, where the prediction uses the best linear predictor of r given $\hat{M}\hat{r}$, evaluated at the true hyperparameters (μ, σ^2). As noted byBrown [1990], shrinkage estimation of a high-dimensional nuisance component can reduce risk for a low-dimensional target in linear models even when $p = 1$.

Surprisingly, removing the predictable component of the first-order estimation error Λr can be valuable even when the model is properly specified. To see this, note that under correct specification the oracle-adjusted estimator satisfies $\sqrt{n}(\theta^* - \theta_0) = \Lambda(I - \Pi_V)\varepsilon + o_p(1)$. The asymptotic variance of this estimator is $\Lambda(I - \Pi_V)V(I - \Pi_V)'\Lambda'$, which equals $(G'V^{-1}G)^{-1}$ — the semiparametric efficiency bound — for any initial weighting matrix W. Thus, the BLP correction achieves full efficiency as a byproduct of predicting the estimation error ε.

In general, θ^* weakly improves on the oracle bias-corrected estimator $\hat{\theta} - \frac{\mu}{\sqrt{n}} \cdot \hat{\Lambda} 1_m$ and can improve on the MSE of θ^*_{BC} even when $p = 1$. The scope for improvement depends on $\|M\Sigma\Lambda'\|$, which itself depends on the interaction between the weighting matrix W and the moment covariance Σ. The following lemma formally characterizes when the linear shrinkage oracle strictly improves upon standard bias correction.

Lemma 3. *Defining the MSE matrix $R(\hat{\theta}) = E[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)']$, the oracle shrinkage estimator θ^* achieves weakly lower risk than the oracle bias-corrected estimator θ^*_{BC}:*

$$R(\theta^*_{BC}) - R(\theta^*) = \frac{1}{n}\Lambda\Sigma M' \left(M\Sigma M'\right)^+ M\Sigma\Lambda' + o(1/n),$$

This matrix is positive semi-definite, as it equals $(1/n) \cdot A'(M\Sigma M')^+A$, where $A = M\Sigma\Lambda'$ and $(M\Sigma M')^+$ is positive semi-definite. Since Assumption 2(vi) ensures $(M\Sigma M')^+$ is positive definite on the range of M, this $p \times p$ matrix is positive semidefinite and nonzero if and only if $M\Sigma\Lambda' \neq 0$. When $p = 1$, nonzeroness and positive definiteness coincide.

Equivalent characterizations:

i) When $p = 1$, $M\Sigma\Lambda' \neq 0$ if and only if ΣWG is **not** proportional to G.

ii) When $p > 1$, $M\Sigma\Lambda' \neq 0$ if and only if there is no $p \times p$ matrix C such that $\Sigma WG = GC$.

Proof. See appendix. □

To illustrate, consider efficient GMM ($W = V^{-1}$). When $\sigma^2 = 0$, $\Sigma WG = G \in \text{col}(G)$ and EB coincides with BC. When $\sigma^2 > 0$, $\Sigma WG = (I + \sigma^2 V^{-1})G$, and $\Sigma WG \in \text{col}(G)$ requires $V^{-1}G \in \text{col}(G)$, a knife-edge condition that typically fails under heteroskedasticity. Thus, under efficient GMM, the shrinkage estimator typically improves on bias correction whenever misspecification is present.

5.2 Plug-in estimator

To build feasible analogues of $L\left[r \mid Mr\right]$, one needs estimators of the signal and noise variances of the moment conditions. The plug-in estimator $\hat{S} = \widehat{\sigma^2} I_m$ captures the specification error variance, while $\hat{\Sigma} = \hat{V} + \hat{S}$ estimates the total variance of r. The plug-in analogue of $L\left[r \mid Mr\right]$ predicts r using $\hat{r}$

$$\hat{L}\left[r \mid \hat{r}; \hat{\mu}, \widehat{\sigma^2}\right] := \hat{\mu} 1_m + \hat{\Sigma}\hat{M}' \left(\hat{M}\hat{\Sigma}\hat{M}'\right)^+ \left(\hat{r} - \hat{\mu}\hat{M}1_m\right).$$

Because $\hat{M}$ is an oblique projection matrix of rank $m-p$, the estimated variance matrix $\hat{M}\hat{\Sigma}\hat{M}'$ is singular, necessitating the use of the Moore-Penrose pseudoinverse.

Using this projection to remove the predicted first-order estimation error yields the Empirical Bayes shrinkage estimator:

$$\hat{\theta}_{EB} := \hat{\theta} - \frac{1}{\sqrt{n}}\hat{\Lambda}\hat{L}\left[r \mid \hat{r}\right].$$

The following Proposition establishes the approximation error of both the projection estimator $\hat{L}[r \mid \hat{r}]$ and the downstream Empirical Bayes estimator $\hat{\theta}_{EB}$.

Proposition 2. *Under Assumptions 1 and 2, the feasible predictor satisfies:*

$$\frac{1}{\sqrt{m}}\|\hat{L}\left[r \mid \hat{r}; \hat{\mu}, \widehat{\sigma^2}\right] - L\left[r \mid Mr; \mu, \sigma^2\right]\| = O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right).$$

Furthermore, the Empirical Bayes estimator satisfies:

$$\hat{\theta}_{EB} - \theta^* = O_p\left(\frac{1}{\sqrt{mn}}\right) + O_p\left(\frac{\sqrt{m}}{n}\right).$$

It follows that the Empirical Bayes estimator satisfies:

$$\hat{\theta}_{EB} - \theta_0 = O_p\left(\frac{1}{\sqrt{mn}}\right) + O_p\left(\frac{\sqrt{m}}{n}\right).$$

Proof. See appendix. □

Contrasting this with Proposition 1, the feasible EB estimator converges at the same rate as the feasible BC estimator. Corollary 1 below establishes the EB CLT. Lemma 3 shows that the oracle EB estimator achieves weakly lower risk than the oracle BC estimator, with a strict improvement when the conditions of Lemma 3 are satisfied.

Remark 4. When $\left\|\hat{M}1_m\right\|^2 \approx 0$, the mean μ is unidentified and the BC estimator coincides with GMM. The EB estimator remains operational by setting $\hat{\mu} = 0$, yielding a pure shrinkage correction driven by $\hat{\sigma}^2$ alone.

6 Distribution theory

The rate results of Propositions 1 and 2 can be sharpened to distributional convergence under the following added regularity condition.

Assumption 3. *The specification errors have uniformly bounded moments:* $\sup_m E[|\xi_j|^k] < \infty$ *for every positive integer* k.

Assumption 3 combines with Assumption 2 to upgrade the probability convergence of quadratic forms in ξ (guaranteed by Assumption 2(vii)) to convergence of all moments, via uniform integrability. The bounded moments condition is satisfied whenever ξ is *i.i.d.* with a finite moment generating function, or rotation invariant with bounded radial moments. Note that a Lindeberg-type negligibility condition on the rows of G follows directly from the bounded-entry condition in Assumption 2(iv) with p fixed.

The distributional convergence results below rely on a moment-matching approach to establishing an exchangeable central limit theorem (CLT): normality is established by verifying that all moments of the standardized estimator converge to Gaussian moments. Assumption 3 guarantees that all moments of the standardized statistic exist and are uniformly bounded, enabling verification of convergence to Gaussian moments.

A key subtlety in establishing distributional convergence is that the oracle variance $\Lambda\Sigma\Lambda' = O(1/m)$ degenerates as $m \to \infty$. The CLT below uses an asymptotic variance that accounts for the $\hat{\mu}$-correction and coincides with $\Lambda\Sigma\Lambda'$ only when $\Lambda 1_m = 0$.

Define the weight vector

$$w := \frac{M'M1_m}{\|M1_m\|^2}$$

which satisfies $w'1_m = 1$ and characterizes the oracle mean estimator $\hat{\mu}_{\text{oracle}} - \mu = w'\eta$, where $\eta := \sigma\xi + \varepsilon$ is the centered reduced form. The following theorem provides a CLT for non-i.i.d. exchangeable data that characterizes the asymptotic behavior of the bias-corrected estimator.

Theorem 1. *Under Assumptions 1–3, conditional on (υ, ς), the feasible bias-corrected estimator satisfies:*

$$V_{BC}^{-1/2}\sqrt{n}\left(\hat{\theta}_{BC} - \theta_0\right) \xrightarrow{d} N\left(0, I_p\right)$$

where $V_{BC} = B\Sigma B'$ with $B := \Lambda\left(I_m - 1_m w'\right)$.

Proof. See appendix. □

The effective sensitivity matrix $B = \Lambda\left(I_m - 1_m w'\right)$ satisfies $B1_m = 0$: the mean specification error μ has been partialled out by the $\hat{\mu}$-correction, analogous to the Frisch–Waugh–Lovell theorem. The asymptotic variance $V_{BC} = B\Sigma B'$ differs from the naive variance $\Lambda\Sigma\Lambda'$ by correction terms that are $o(1/m)$ under Assumption 2(iv), since all correction terms involve $\Lambda 1_m = -(G'WG)^{-1}G'W1_m = o(1)$ by Assumption 2(iv). Thus, V_{BC} and $\Lambda\Sigma\Lambda'$ are asymptotically equivalent, though the correction from estimating μ can affect the width of confidence intervals in finite samples where $\Lambda 1_m$ is non-negligible.

Corollary 1. *Under Assumptions 1–3, conditional on (υ, ς), the feasible Empirical Bayes estimator satisfies:*

$$V_{EB}^{-1/2}\sqrt{n}\left(\hat{\theta}_{EB} - \theta_0\right) \xrightarrow{d} N\left(0, I_p\right)$$

where $V_{EB} = B_{EB}\Sigma B_{EB}'$ with $B_{EB} := \Lambda\left(I_m - \Pi\right)\left(I_m - 1_m w'\right)$.

Proof. See appendix. □

Here $\Pi = \Sigma M'(M\Sigma M')^{+}M$ is the BLP coefficient for r given Mr, so that $I_m - \Pi$ removes the linearly predictable component of the estimation error. The $\hat{\sigma}^2$-estimation error is second-order for the empirical Bayes CLT: it contributes $O_p(1/m)$ to $\sqrt{n}\left(\hat{\theta}_{EB} - \theta_0\right)$, which is $o_p(1/\sqrt{m})$ and negligible after standardization. Consequently, the asymptotic variance has the same structural form as the bias-corrected case, with Λ replaced by $A = \Lambda\,(I_m - \Pi)$.

Lemma 3 establishes that the oracle shrinkage estimator θ^* achieves weakly lower risk than the oracle bias-corrected estimator θ^*_{BC}, with strict improvement when $M\Sigma\Lambda' \neq 0$. For the feasible estimators, the asymptotic variances V_{BC} and V_{EB} differ from their oracle counterparts by centering corrections of order $O(1/m)$ arising from estimation of μ. Since the oracle risk gap is itself $O(1/m)$, these corrections are of the same order, and a formal matrix ordering $V_{BC} \geq V_{EB}$ does not follow from the oracle result alone. Nevertheless, the oracle comparison characterizes the leading-term risk structure, and the simulations in Section 7 confirm the risk ordering holds in practice. Under correct specification, V_{EB} reduces to $\left(G'V^{-1}G\right)^{-1} + o\,(1/m)$. Therefore, $\hat{\theta}_{EB}$ is asymptotically equivalent to efficient GMM regardless of the initial weighting matrix W. The two estimators differ numerically because $\hat{\theta}_{EB}$ applies a linearized correction rather than re-optimizing the GMM objective.

Feasible inference follows from plug-in estimation of the asymptotic variances. Define $\hat{B} = \hat{\Lambda}\,(I_m - 1_m\hat{w}')$ with $\hat{w} = \hat{M}'\hat{M}1_m / \left\|\hat{M}1_m\right\|^2$, and $\hat{B}_{EB} = \hat{\Lambda}\left(I_m - \hat{\Pi}\right)(I_m - 1_m\hat{w}')$. The feasible variance estimators $\hat{V}_{BC} = \hat{B}\hat{\Sigma}\hat{B}'$ and $\hat{V}_{EB} = \hat{B}_{EB}\hat{\Sigma}\hat{B}'_{EB}$, with $\hat{\Sigma} = \hat{\sigma}^2 I_m + \hat{V}$, are consistent under the rates in Assumption 2(iii). Asymptotic $100 \cdot (1-\gamma)\%$ confidence intervals for the k-th component of θ_0 take the form $\hat{\theta}_{BC,k} \pm z_{1-\gamma/2}\sqrt{\hat{V}_{BC,kk}/n}$, and analogously for EB.

7 Monte Carlo Evidence

This section presents Monte Carlo evidence on the finite-sample performance of the estimators developed above. The simulations are based on the overidentified instrumental variables model of Example 4.

7.1 Simulation design

The data generating process (DGP) is

$$
\begin{aligned}
Y_i &= \alpha_0 + \beta_0 D_i + \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{m-1} b_j Z_{ij} + b_m \right) + e_i, \\
D_i &= \mathbf{1}\{Z_i'\pi + \nu_i > 0\}, \quad e_i = \sigma_e \left(\rho \nu_i + \sqrt{1-\rho^2}\eta_i \right), \\
b &\sim t_8(\bar{\mu}\mathbf{1}_m, \bar{\sigma}^2 I_m), \quad \mathrm{Var}(b_j) = \bar{\sigma}^2, \quad \nu_i \overset{iid}{\sim} N(0,1), \quad \eta_i \overset{iid}{\sim} N(0,1), \\
&\nu_i \perp \eta_i \perp Z_i \perp b.
\end{aligned}
$$

The parameter $\rho = \mathrm{Corr}(e_i, \nu_i)$ governs the degree of endogeneity of D_i, while the scale parameter σ_e controls the variance of the structural error. The vector $Z_i = (Z_{i1}, \ldots, Z_{im-1})'$ is comprised of mutually exclusive binary instruments $Z_{ij} \in \{0,1\}$ obeying $Z_{ij}Z_{i\ell} = 0$ for $j \neq \ell$. We assume

$$
\begin{pmatrix} 1 - \sum_{j=1}^{m-1} Z_{ij} \\ Z_i \end{pmatrix} \sim \text{Multinomial}\left(1, \begin{pmatrix} 1 - \sum_{j=1}^{m-1} q_j \\ q \end{pmatrix}\right), \quad q_j > 0 \text{ for } j = 1, \ldots, m-1.
$$

This design can be thought of as a setting where a single fundamental instrument $\sum_{j=1}^{m-1} Z_{ij}$ has been fully interacted with $m-1$ group indicators. The term b_m in the DGP plays the role of a direct effect of belonging to the omitted category, which contaminates the case where the fundamental instrument equals zero. The complete instrument set is $(1, Z_i')' = (1, Z_{i1}, \ldots, Z_{im-1})'$, giving $m > 3$ moment conditions for the 2×1 parameter vector $\theta_0 = (\alpha_0, \beta_0)$.

The vector $b = (b_1, \ldots, b_m)$ captures excludability violations that follow a multivariate t distribution with 8 degrees of freedom, location $\bar{\mu}\mathbf{1}_m$, and covariance $\bar{\sigma}^2 I_m$ (equivalently, scale matrix $\frac{\nu-2}{\nu}\bar{\sigma}^2 I_m = \frac{3}{4}\bar{\sigma}^2 I_m$ in the standard $t_\nu(\mu, \Sigma)$ parameterization). Hence, the components of b are exchangeable but not *i.i.d.*: the isotropic covariance ensures invariance to permutations, while the common mixing variable induces dependence across components. The specification error admits the scale-mixture representation $b = \bar{\mu}\mathbf{1}_m + \varsigma\bar{\sigma}\xi$, where $\varsigma \sim \sqrt{6/\chi^2_8}$ and $\xi \sim \mathcal{N}(0, I_m)$ are independent. Thus, conditioning on the random scale parameter ς yields conditionally independent specification errors with identical means $\mu = \bar{\mu}$ and variances $\sigma^2 = \varsigma^2\bar{\sigma}^2$. Importantly, the vector b is scaled by $n^{-1/2}$, which ensures that GMM remains consistent for the true parameter θ_0 rather than a pseudo-true parameter.

Define the second moment matrix

$$Q := E_n\left[\begin{pmatrix} 1 \\ Z_i \end{pmatrix}(1, Z_i')\right] = \begin{bmatrix} 1 & q' \\ q & diag\,(q) \end{bmatrix}.$$

Per Example 4, I work with the scaled instrumental variable moment conditions

$$g\,(\theta) = Q^{-1}E_n\left[(Y_i - \alpha - \beta D_i)\begin{pmatrix} 1 \\ Z_i \end{pmatrix}\right].$$

To verify that Assumption 1 holds, note that

$$E_n\left[(Y_i - \alpha_0 - \beta_0 D_i)\begin{pmatrix} 1 \\ Z_i \end{pmatrix}\right] = \left(\sum_{j=1}^{m-1} b_j q_j + b_m, (b_1 + b_m)\,q_1, \ldots, (b_{m-1} + b_m)\,q_{m-1}\right)' / \sqrt{n} + o\left(n^{-1/2}\right).$$

Therefore, premultiplying by Q^{-1} yields $\sqrt{n}g\,(\theta_0) = (b_m, b_1, \ldots, b_{m-1}) + o\,(1)$, confirming that the moment errors are exchangeable.

Letting $\Xi = diag\,(q) - qq'$ and $\delta = Cov_n\,(D_i, Z_i)$, the Jacobian of this model is

$$G = -\begin{pmatrix} 1 & E_n\,[D_i] - q'\Xi^{-1}\delta \\ 0_{m-1} & \Xi^{-1}\delta \end{pmatrix}.$$

In general, 1_m will not be in the column space of G. Hence, μ will be identified.

Both π and q are deterministic functions of m and are held fixed across all Monte Carlo replications. Group membership probabilities are set to quantiles of a log-normal distribution: $q_j^{\text{raw}} = \exp\big(\sigma_q\,\Phi^{-1}(u_j)\big)$ where $u_j = (j - 0.5)/(m - 1)$ and $\sigma_q = 0.83$, then normalized so that $\sum_{j=1}^{m-1} q_j = 0.9$. This gives a spread $\max_j q_j / \min_j q_j \approx 40{:}1$ at $m = 40$, with the shape preserved as m grows. First-stage coefficients use the same quantile scheme: $\pi_j^{\text{raw}} = \exp\big(\sigma_\pi\,\Phi^{-1}(u_j)\big)$ with $\sigma_\pi = 0.67$, giving a roughly 20:1 spread. Then the raw values are rescaled so that $\|\pi\| = 3\sqrt{(m-1)/39}$, which preserves average instrument strength as m grows. To break correlation with group sizes, π values are assigned in an interleaved pattern: odd-indexed groups receive π values in ascending order and even-indexed groups in descending order.

In the baseline parameterization, I set $\alpha_0 = 1$, $\beta_0 = 1$, $\rho = 0.1$, $\bar{\mu} = 8$, $\bar{\sigma} = 8$, and $\sigma_e^2 = 1$. Recall that b is scaled by $n^{-1/2}$. Hence, the coefficients capturing excludability violations have mean $8/\sqrt{n}$ and variance $64/n$.

The asymptotic variance of the scaled sample moment conditions $\hat{g}\,(\theta_0)$ is $V = Q^{-1}\sigma_e^2$. Therefore the

optimal weighting matrix under correct specification is $W = Q$, which ensures Assumption 2(ii) is without loss of generality. Letting $\hat{q} := n^{-1} \sum_{i=1}^{n} Z_i$, I use the plug-in weighting matrix

$$\hat{W} = \begin{bmatrix} 1 & \hat{q}' \\ \hat{q} & diag(\hat{q}) \end{bmatrix}.$$

By invariance, weighting the scaled moment conditions by $\hat{W}$ is numerically equivalent to running efficient GMM on the original unscaled moment conditions with weighting matrix $\hat{W}^{-1}$, thus providing a natural benchmark for standard GMM estimation predicated on correct specification.

7.2 Baseline findings

Table 1 reports the bias, standard deviation, and root mean squared error (RMSE) of each estimator in a first simulation design with $n = 10{,}000$ observations and $m = 40$ instruments. All results report estimation performance for the slope coefficient β and are averaged across 1,000 Monte Carlo replications. The feasible estimators use the hyperparameter estimates $\hat{\mu}$ and $\hat{\sigma}^2$, while the oracle estimators use the true (conditional) hyperparameters (μ, σ^2) together with the sample matrices $\hat{G}$, $\hat{V}$, and $\hat{W}$ to bias-correct GMM.

Let Λ_β denote the row of the population sensitivity matrix corresponding to the parameter β. The quantity $\|\Lambda_\beta \Sigma M'\| = 89.5$ is nonzero in this design, which (via Lemma 3) suggests that the oracle shrinkage estimator should strictly improve upon oracle bias correction. The mean simulated J statistic is 94.33, which exceeds the 5% critical value of 53.38. Faced with this DGP, the overidentification test will tend to reject correct specification.

Table 1: Estimator Performance ($m = 40$, $n = 10,000$)

Estimator	Bias	Std	RMSE	Rej.
GMM	0.219	0.180	0.284	0.373
BC	0.041	0.239	0.243	0.064
EB	0.041	0.232	0.236	0.068
Oracle BC	0.023	0.178	0.180	0.046
Oracle EB	0.017	0.170	0.171	0.058

Notes: Bias, standard deviation, and RMSE of $\hat{\beta}$ across 1,000 replications, $n = 10{,}000$, $m = 40$. Feasible estimators use the hyperparameter estimates described in Section 4. "Oracle" estimators use the true (μ, σ^2) with sample matrices $\hat{G}$, $\hat{V}$, $\hat{W}$. Mean simulated $J = 94.33$ with 38 degrees of freedom. Population $\|\Lambda_\beta \Sigma M'\| = 89.5$ computed via 2 million simulations. Rej. is the rejection rate of a two-sided Wald test of the null hypothesis that $\beta_0 = 1$ at the 5% nominal level.

The sizeable specification errors present in this DGP lead the uncorrected GMM estimator to exhibit a bias of 0.219 that accounts for the bulk of its RMSE of 0.284. The bias-corrected estimator $\hat{\beta}_{BC}$ reduces the bias to 0.041 and achieves an RMSE of 0.243, 14% below GMM. The feasible shrinkage estimator $\hat{\beta}_{EB}$ does

slightly better: RMSE 0.236, 17% below GMM. The oracle EB estimator reaches an RMSE of 0.171, 40% below GMM, with the oracle OEB/OBC RMSE ratio at $0.171/0.180 = 0.95$, confirming a 5% RMSE gain from shrinkage at the oracle level.

The "Rej." column of Table 1 reports the rejection rate of a two-sided Wald test of the null hypothesis that $\beta_0 = 1$ at the 5% nominal level. Thus, rejections correspond here to type I errors. The standard GMM Wald test over-rejects dramatically, with a rejection rate of 37.3%. This overrejection stems both from the estimator's bias and use of the conventional sandwich variance estimator that ignores misspecification. The feasible BC and EB tests over-reject moderately at roughly 6.5%, reflecting the finite-sample cost of estimating the hyperparameters μ and σ^2. The oracle BC and EB rejection rates are based on the following infeasible variance matrices

$$V_{OBC} = \hat{\Lambda}\left(\hat{V} + \sigma^2 I_m\right)\hat{\Lambda}', \quad V_{OEB} = \hat{\Lambda}\left(I_m - \hat{\Pi}\right)\left(\hat{V} + \sigma^2 I_m\right)\left(I_m - \hat{\Pi}\right)'\hat{\Lambda}',$$

which use the sample sensitivity matrix $\hat{\Lambda}$, the sample moment-covariance matrix $\hat{V}$, and the true σ^2. The oracle BC and EB rejection rates are 4.6% and 5.8% respectively, both close to the nominal level, confirming the asymptotic validity of the distributional approximations derived in Section 6.

7.3 Altering the misspecification hyperparameters

Table 2 examines how the relative performance of the estimators, as measured by RMSE, varies with the dispersion $\bar{\sigma}$ of the specification errors while holding $m = 40$, $\bar{\mu} = 8$, and all other parameters fixed. The first row reports the mean simulated J-statistic, which rises from 40.3 at $\bar{\sigma} = 0$ to 244.5 at $\bar{\sigma} = 16$.

Table 2: RMSE by $\bar{\sigma}$ ($\bar{\mu} = 8$, $n = 10{,}000$, $m = 40$)

	$\bar{\sigma} = 0$	$\bar{\sigma} = 2$	$\bar{\sigma} = 4$	$\bar{\sigma} = 8$	$\bar{\sigma} = 16$
J-stat	40.3	43.6	54.1	94.3	244.5
GMM	0.264	0.265	0.273	0.284	0.350
BC	0.180	0.185	0.203	0.243	0.380
EB	0.182	0.187	0.203	0.236	0.345
Oracle BC	0.144	0.147	0.158	0.180	0.275
Oracle EB	0.146	0.148	0.156	0.171	0.224
OEB/OBC	1.02	1.01	0.99	0.95	0.81

Notes: RMSE of $\hat{\beta}$ across 1,000 Monte Carlo replications with $m = 40$, $n = 10{,}000$, and $\bar{\mu} = 8$ held fixed. The column $\bar{\sigma} = 8$ is the baseline. First row reports the mean simulated J statistic. "OEB/OBC" gives the RMSE ratio; values below 1 favor the shrinkage estimator.

At $\bar{\sigma} = 0$ (homogeneous specification error), OEB RMSE (0.146) slightly exceeds OBC (0.144), and the feasible estimators are close (EB 0.182 vs. BC 0.180). As $\bar{\sigma}$ grows, the signal from the specification

errors strengthens relative to sampling noise, making shrinkage increasingly beneficial: the OEB/OBC ratio declines from 1.02 to 0.95 at $\bar{\sigma} = 8$ and reaches 0.81 at $\bar{\sigma} = 16$. GMM RMSE increases modestly from 0.264 to 0.350 over this range. At $\bar{\sigma} = 16$, EB dominates with an RMSE of 0.345 versus 0.380 for BC, roughly a 9% advantage.

Table 3 complements the preceding analysis by varying the mean specification error $\bar{\mu}$ (which equals μ because the t-distribution is a scale but not location mixture) while holding $\bar{\sigma} = 8$ and all other parameters fixed. Theoretically, the BC, Oracle BC, EB, and Oracle EB estimators are invariant to $\bar{\mu}$ because the bias correction is linear in the mean shift, which cancels exactly; only the uncorrected GMM estimator's RMSE genuinely depends on $\bar{\mu}$. The mild variation visible in the corrected-estimator columns of the table reflects Monte Carlo sampling noise rather than a true dependence on $\bar{\mu}$.

Table 3: RMSE by $\bar{\mu}$ ($\bar{\sigma} = 8$, $n = 10{,}000$, $m = 40$)

	$\bar{\mu} = 0$	$\bar{\mu} = 2$	$\bar{\mu} = 4$	$\bar{\mu} = 8$	$\bar{\mu} = 16$
J-stat	89.2	93.3	92.6	94.3	102.0
GMM	0.191	0.202	0.221	0.284	0.459
BC	0.251	0.248	0.248	0.243	0.251
EB	0.242	0.239	0.241	0.236	0.243
Oracle BC	0.191	0.190	0.185	0.180	0.184
Oracle EB	0.177	0.176	0.173	0.171	0.172
OEB/OBC	0.93	0.92	0.94	0.95	0.94

Notes: RMSE of $\hat{\beta}$ across 1,000 Monte Carlo replications with $m = 40$, $n = 10{,}000$, and $\bar{\sigma} = 8$ held fixed. First row reports the mean simulated J statistic. "OEB/OBC" gives RMSE ratios; values below 1 favor the shrinkage estimator.

At $\bar{\mu} = 0$ (no mean specification error), the corrections add noise when there is no systematic bias to correct, leading GMM (0.191) to beat BC (0.251) and EB (0.242) by a wide margin. At $\bar{\mu} = 8$, the corrections dominate: EB achieves 0.236 versus 0.284 for GMM. At $\bar{\mu} = 16$, GMM degrades to 0.459 while BC and EB stay around 0.24--0.25—a massive improvement. The OEB/OBC ratio ranges from 0.92 to 0.95 across the $\bar{\mu}$ grid, confirming a stable advantage of shrinkage over pure bias correction.

7.4 Asymptotic behavior

To assess the rate predictions of Proposition 2, I vary n and m jointly, setting m equal to $n^{0.4}$ rounded to the nearest integer. This choice ensures that $m^2/n = n^{-0.2} \to 0$ as required by Assumption 2. Table 4 reports the bias and RMSE of all five estimators across seven sample sizes from $n = 5{,}000$ to $n = 500{,}000$.

Table 4: Bias and RMSE as Functions of n with $m \approx n^{0.4}$

			Bias					RMSE				
n	m	$\|\Lambda_\beta \Sigma M'\|$	GMM	BC	EB	OBC	OEB	GMM	BC	EB	OBC	OEB
5,000	30	102.04	0.299	0.061	0.052	0.045	0.029	0.418	0.380	0.360	0.291	0.261
10,000	40	89.52	0.219	0.041	0.038	0.023	0.014	0.284	0.243	0.236	0.180	0.170
20,000	53	78.76	0.164	0.027	0.024	0.016	0.011	0.207	0.170	0.166	0.127	0.122
50,000	76	65.08	0.110	0.016	0.014	0.011	0.008	0.135	0.109	0.107	0.079	0.076
100,000	100	56.76	0.085	0.019	0.018	0.013	0.011	0.101	0.077	0.076	0.056	0.055
200,000	132	50.69	0.060	0.011	0.010	0.007	0.006	0.071	0.053	0.053	0.039	0.039
500,000	190	43.88	0.038	0.007	0.007	0.005	0.004	0.046	0.033	0.033	0.025	0.025

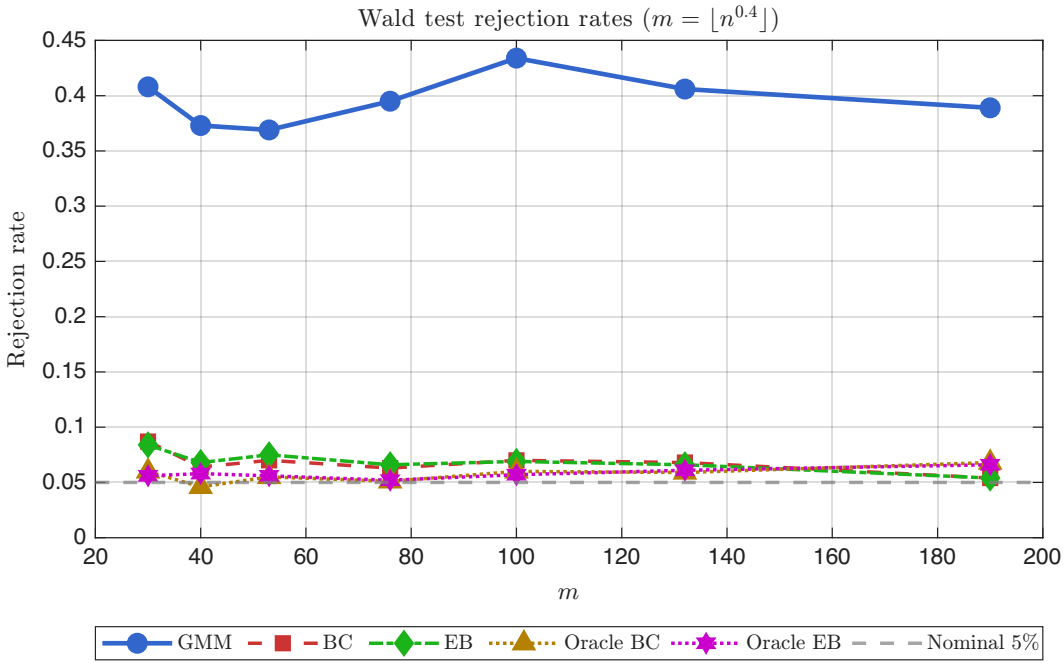
Notes: Bias and RMSE of $\hat{\beta}$ across 1,000 replications per grid point. Population $\|\Lambda_\beta \Sigma M'\|$ computed via 2 million simulations.

In accord with the theoretical prediction that both estimators are consistent as $m^2/n \to 0$, the BC and EB biases decline toward zero as sample size grows. At $n = 500{,}000$ ($m = 190$), the oracle BC bias is 0.005 and the oracle EB bias is 0.004, comparable to their feasible counterparts (0.007 and 0.007 respectively). The bias of GMM also declines with n but remains substantially larger than the corrected estimators at every sample size. BC and EB substantially reduce GMM bias throughout the grid.

GMM has the highest RMSE at every sample size. EB and BC yield marked improvements on the RMSE of GMM; however, the RMSE of the two approaches converges as n grows large. Oracle EB has the lowest RMSE throughout but the RMSE of OEB and OBC approaches also converge once $n = 500{,}000$. Both forms of convergence are attributable to the tendency of $\|\Lambda_\beta \Sigma M'\|$ to decline with n—it falls from 102 to 44 as m grows across the grid—which reflects the growing number of moments over which the shrinkage operates. Ultimately, the absolute scale of these RMSE differences dissipates as all estimators become consistent — a feature of the local misspecification framework — though the relative performance gap and inference distortions persist, as the following figure makes clear.

Figure 1 plots rejection rates across the Table 4 grid. GMM's bias leads to over-rejection at all sample sizes, hovering between 37% and 43% type I error rates as m grows large. In contrast, both the BC and EB rejection rates approach the nominal 5% level as m grows large. The oracle BC and EB rejection rates hover in the range 4-7% across all m, with both OBC (6.8%) and OEB (6.7%) exhibiting mild over-rejection at $m = 190$.

Figure 1: Wald test rejection rates vs. m with $m \approx n^{0.4}$ (nominal 5%, 1,000 replications per grid point).



Notes: Figure depicts rejection rate of Wald test that $\hat{\beta} = \beta_0 = 1$ for each estimator.

8 Empirical Application: Returns to Schooling

In this section, I revisit the work of Angrist and Krueger [1991], who used quarter-of-birth (QOB) indicators as instruments for years of education in a log weekly wage equation to estimate the returns to schooling. Table VII of their paper reports TSLS estimates based on interactions between QOB and state-of-birth (SOB) as excluded instruments. The same specification also excludes interactions between QOB and year-of-birth (YOB), yielding 180 excluded instruments in total.

I will simplify this instrument set in two respects with the dual aims of ensuring exchangeability and avoiding potential biases that can arise under many instrument asymptotics. First, I collapse the three QOB indicators into a single $\mathbf{1}\{\text{QOB} > 1\}$ indicator for being born in one of the last three quarters, interacted with SOB. Angrist and Krueger [1991] work with the same binarized version of QOB in their Table III, which they show yields similar results to specifications leveraging the complete set of QOB indicators as instruments via TSLS. Second, I drop the season of birth interactions with YOB. This yields 51 instruments (one per state) that are plausibly exchangeable.

A key identifying assumption of Angrist and Krueger [1991]'s study is that season of birth is correlated with earnings only through its causal effect on educational attainment. This claim was questioned early on by Bound et al. [1995], who considered the potential effects of small violations of QOB exogeneity. Later

work by Buckles and Hungerman [2013] provided evidence that economically disadvantaged families are more likely to have children in the first quarter of the year, suggesting the average earnings of children born in later quarters would be higher even if their years of schooling were equalized. Since Angrist and Krueger [1991] establish that individuals born in later quarters tend to complete more schooling, the family background differences highlighted by Buckles and Hungerman [2013] may lead QOB-based IV estimates to overstate the returns to schooling when instruments have a positive first stage. In what follows, I investigate whether the BC and EB estimators reach similar conclusions about the likely bias in TSLS.

8.1 Estimation sample and moment conditions

The estimation sample consists of men from the 1940--1949 birth cohorts in the 1980 Census 5% Public Use Micro Sample (PUMS), yielding $n = 486{,}926$ observations. The instruments are $m = 51$ mutually exclusive indicators of the form $Z_{ij} = \mathbf{1}\{\text{QOB}_i > 1, \text{SOB}_i = s_j\}$ for each state s_j, $j = 1, \ldots, 51$. Each instrument flags whether individual i was born in a non-Q1 quarter in a particular state. Note that $\sqrt{n} \approx 698$ and $m^2/n \approx 0.005$, which suggests the asymptotic regime of Assumption 2(i) is easily satisfied.

I consider four specifications of controls X_i mirroring those entertained by Angrist and Krueger [1991]: (i) SOB dummies only, (ii) SOB dummies with age (in quarters) and age-squared, (iii) SOB dummies with individual-level covariates (indicators for race, marital status, SMSA residence, and eight census region dummies), and (iv) SOB dummies with all covariates plus age and age-squared. Rather than include a constant, I always use 51 state-of-birth dummies that sum to one.

The endogenous variable is completed years of schooling S_i. The scalar parameter of interest is θ, which measures the returns to an additional year of schooling. Since the controls are nuisance parameters, I partial them out. Thus, in all four specifications the degree of overidentification is $m - p = 50$. As the simulations of the previous section demonstrate, this level of overidentification can yield sufficiently accurate hyperparameter estimates for corrected estimators to generate substantial improvements over GMM.

Let Y_i denote log weekly earnings and let $\tilde{Y}_i$ and $\tilde{S}_i$ denote the residuals from OLS regressions of Y_i and S_i, respectively, on X_i. I work with moment conditions of the form

$$\hat{g}(\theta) = \hat{Q}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i - \theta\tilde{S}_i\right)Z_i\right), \quad \hat{Q} = \operatorname{diag}(\hat{q}), \quad \hat{q} = \frac{1}{n}\sum_i Z_i.$$

Because the Z_{ij} are mutually exclusive binary indicators, the scaled moment conditions reduce to cell means: $\hat{g}_j(\theta) = \bar{Y}_j - \theta\bar{S}_j$, where overbars denote cell means of residualized variables. Evaluated at the true parameter θ_0, the population moment $g_j(\theta_0)$ gives the direct effect of being born in a non-Q1 quarter in state s_j relative to that state's average. Exchangeability of these direct effects across states is plausible ex-ante. For example,

the family composition mechanism of Buckles and Hungerman [2013] could operate similarly in each state. I develop some diagnostics for empirically scrutinizing the exchangeability assumption below.

The sample Jacobian is the column vector $\hat{G} = -(\bar{S}_1, \ldots, \bar{S}_{51})'$. Since non-Q1 births have more education, it is reasonable to expect $\bar{S}_j > 0$ for each j and $\hat{G} < 0$. The vector 1_m is in the column space of $\hat{G}$ only if all first stages are equal across states. Fortunately, cross-state variation in compulsory schooling laws yields substantial variation in $\bar{S}_j$ across states, which was Angrist and Krueger [1991]'s motivation for studying SOB interactions.

8.2 OLS and TSLS Results

Table 5 reports OLS and TSLS estimates for the non-Q1 × SOB instrument design. Without age controls, the TSLS estimates of the return to education are around 4%, below the OLS estimates of approximately 5%. Adding age controls raises the TSLS estimates sharply, to 12--14%, well above OLS. This sensitivity to the inclusion of age reflects the mechanical correlation between age and QOB: non-Q1 individuals are younger on average. Rosenzweig and Wolpin [2000] argued that this mechanical dependence led to excludability difficulties as younger workers have less labor market experience and therefore lower potential wages. The sizable increase in the TSLS estimates after controlling for the age quadratic is consistent with this view.

Table 5: TSLS Estimates: Returns to Education with nonQ1 × SOB Instruments

	SOB	SOB + age	SOB + cov	SOB + cov + age
OLS	0.0535	0.0555	0.0495	0.0513
	(0.0004)	(0.0004)	(0.0003)	(0.0003)
TSLS	0.0424	0.1413	0.0360	0.1240
	(0.0188)	(0.0246)	(0.0193)	(0.0251)
J [d.f.]	67.32 [50]	46.64 [50]	58.66 [50]	45.65 [50]
p-value	0.052	0.609	0.188	0.649

Notes: Sample is 486,926 men from the 1940--1949 birth cohorts in the 1980 Census 5% PUMS. Instruments are 51 $\mathbf{1}\{\text{QOB} > 1\}$ × SOB interactions. Heteroskedasticity-robust standard errors in parentheses. J-statistic degrees of freedom in brackets.

Unlike in our Monte Carlo simulations, these data exhibit minimal evidence of misspecification. The J-test borderline rejects correct specification in the two specifications without age controls ($p = 0.052$ and $p = 0.188$) but fails to reject in both specifications with age controls ($p = 0.609$ and $p = 0.649$). I will show that misspecification-aware point estimates and standard errors nonetheless offer additional insight in this environment.

To anchor the subsequent misspecification-aware analysis to these TSLS specifications, I employ the GMM weighting matrix $\hat{W} = \hat{Q}\tilde{Q}^{-1}\hat{Q}$, where $\tilde{Q} = \frac{1}{n}\sum_i \tilde{Z}_i\tilde{Z}_i'$ and $\tilde{Z}_i$ is the residual from a regression of Z_i against X_i. This weighting, which ensures the GMM point estimates match the TSLS estimates exactly,

is efficient under correct specification and homoscedasticity of the earnings errors. I estimate moment uncertainty with the heteroscedasticity-robust sandwich estimator $\hat{V}$, computed from the outer product of moment contributions evaluated at the residualized instruments $\tilde{Z}_i$ to account for estimation of the controls.

8.3 Scrutinizing exchangeability

The exchangeability assumption (Assumption 1) requires that the distribution of the specification errors b_j does not depend on observable features of the moment conditions such as first-stage strength or cell size. A simple diagnostic for this restriction can be constructed using the distributional results of Theorem 1.

Let $\tilde{f}_j \in \mathbb{R}^k$ denote a vector of centered moment-level features obeying $\sum_j \tilde{f}_j = 0$ and $\tilde{r} = \hat{r} - \hat{\mu}\hat{M}1_m$ denote the demeaned residuals. Under exchangeability, conditioning on the features gives $E\left[\sum_j b_j \tilde{f}_j \mid \tilde{f}_1, \ldots, \tilde{f}_m\right] = \mu \sum_j \tilde{f}_j = 0$. Hence, the $k \times 1$ vector of sample cross-products $\tilde{F}'\tilde{r}$, where $\tilde{F} = \left(\tilde{f}_1, \ldots, \tilde{f}_m\right)'$, should be close to zero. Under exchangeability, Theorem 1 and the exchangeable CLT arguments used in its proof imply $\tilde{F}'\tilde{r}$ is asymptotically normal with variance $\tilde{F}'M\left(I_m - 1_m w'\right)\Sigma\left(I_m - w1_m'\right)M'\tilde{F}$, where the $I_m - 1_m w'$ projections account for the estimation of μ. This motivates the diagnostic statistic $T = \left(\tilde{F}'\tilde{r}\right)' \Omega^{-1} \left(\tilde{F}'\tilde{r}\right)$ where

$$\Omega = \tilde{F}'\hat{M}\left(I_m - 1_m\hat{w}'\right)\hat{\Sigma}\left(I_m - \hat{w}1_m'\right)\hat{M}'\tilde{F}.$$

Under the null, T is approximately $\chi^2(k)$. Individual features can be assessed via the statistics $\left(\tilde{F}'\tilde{r}\right)_j / \sqrt{\Omega_{jj}}$. Because the distributional theory of Theorem 1 assumes exchangeability, the behavior of T under alternatives is not formally characterized. The diagnostic should therefore be interpreted as a heuristic check rather than a formal hypothesis test.

I take $k = 2$ with features given by centered first-stage strength $\tilde{f}_{1j} = \bar{S}_j - m^{-1}\sum_\ell \bar{S}_\ell$ and centered cell size $\tilde{f}_{2j} = \hat{q}_j - m^{-1}\sum_\ell \hat{q}_\ell$. The first feature tests whether states with stronger compulsory schooling effects exhibit systematically different specification errors, which would arise, for example, if the family composition channel of Buckles and Hungerman [2013] varies with the stringency of compulsory schooling laws. The second tests for the precision dependence discussed in Example 1.

Table 6: Exchangeability Diagnostics

	First-stage t	Cell-size t	$\chi^2(2)$ [p-val]
SOB	−0.77	0.19	0.76 [0.685]
SOB + age	−0.57	0.07	0.37 [0.829]
SOB + cov	−1.06	0.15	1.32 [0.518]
SOB + cov + age	−0.78	−0.03	0.68 [0.713]

Notes: The test statistic $\left(\tilde{F}'\tilde{r}\right)' \Omega^{-1} \left(\tilde{F}'\tilde{r}\right)$ is asymptotically $\chi^2(2)$ under exchangeability, where $\Omega = \tilde{F}'\hat{M}\left(I_m - 1_m\hat{w}'\right)\hat{\Sigma}\left(I_m - \hat{w}1_m'\right)\hat{M}'\tilde{F}$. Features are centered first-stage strength $\tilde{f}_{1j} = \bar{S}_j - m^{-1}\sum_\ell \bar{S}_\ell$ and centered cell size $\tilde{f}_{2j} = \hat{q}_j - m^{-1}\sum_\ell \hat{q}_\ell$. Sample and instruments as in Table 5.

Table 6 reports the resulting test statistics and p-values. In all four specifications, the individual t-statistics are small (the largest is 1.06 in absolute value) and the joint $\chi^2(2)$ test fails to reject at any conventional level ($p > 0.51$). Cell size has essentially no predictive power for the specification errors in any specification. First-stage strength shows a mild negative association across all specifications, but the t-statistics remain well within normal sampling variation. While other sorts of violations could be at play, these results provide no evidence against the exchangeability assumption.

8.4 Misspecification-aware estimates

Table 7 presents estimates of the hyperparameters (μ, σ^2) alongside the GMM, BC, and EB estimates. Again, the GMM point estimates and standard errors coincide with the TSLS estimates reported in Table 5 by construction. The reported J-statistics, however, differ slightly from those in Table 5. The weight matrix choice forces the GMM estimator of θ to equal TSLS, but the J-statistic inverts $\hat{V}$ over only the excluded-instrument block after FWL projection, whereas the TSLS SarganHansen statistic uses the full-system variance including the control block. Under the null of correct specification the two statistics are asymptotically equivalent.

Table 7: Bias-Corrected and Empirical Bayes Estimates: Returns to Education

	SOB	SOB + age	SOB + cov	SOB + cov + age
Hyperparameters				
$\hat{\mu}$	−1.32	1.05	−0.91	1.11
$\hat{\sigma}^2$	4.13	0.00	0.87	0.00
J [d.f.]	66.21 [50]	46.14 [50]	57.76 [50]	45.14 [50]
$\|\hat{\Lambda}\hat{\Sigma}\hat{M}'\|$	22.60	7.75	7.62	7.70
Point estimates				
GMM	0.0424	0.1413	0.0360	0.1240
	(0.0188)	(0.0246)	(0.0193)	(0.0251)
BC	0.0880	0.1120	0.0683	0.0939
	(0.0375)	(0.0317)	(0.0349)	(0.0313)
EB	0.0829	0.1072	0.0634	0.0883
	(0.0370)	(0.0318)	(0.0352)	(0.0314)

Notes: Sample and instruments as in Table 5. All 51 nonQ1 × SOB moment conditions are treated as potentially misspecified. GMM uses weighting matrix $\hat{W} = \hat{Q}\tilde{Q}^{-1}\hat{Q}$ on $\hat{Q}^{-1}$-scaled moments, reproducing the TSLS estimator. Heteroskedasticity-robust standard errors in parentheses.

The estimated mean specification error $\hat{\mu}$ is negative in the specifications without age controls, consistent with the experience channel highlighted by Rosenzweig and Wolpin [2000]: non-Q1 individuals have less potential labor market experience, producing a negative direct effect on earnings. With age controls, $\hat{\mu}$ turns positive, consistent with the family composition mechanism of Buckles and Hungerman [2013]: once

potential experience differences are absorbed, the remaining direct effect of non-Q1 birth reflects better family backgrounds, which raise earnings.

Consistent with the J-statistics falling below their degrees of freedom, the estimated variance $\hat{\sigma}^2$ is zero in both specifications with age controls, indicating no excess dispersion beyond the mean shift once experience differences are absorbed. In the specifications without age controls, the J-statistics exceed their expected values and $\hat{\sigma}^2$ is positive, suggesting some heterogeneity in the direct effects across states.

Without age controls, $\hat{\mu} < 0$ and the BC and EB corrections adjust upward. In the specifications with age controls, $\hat{\mu} > 0$ and the BC and EB corrections adjust downward. In all specifications, the corrected estimates converge toward the 0.06--0.11 range, substantially narrowing the cross-specification dispersion relative to GMM/TSLS. While the corrections consistently shift estimates in the direction of OLS, they overshoot in the specifications without age controls, where the estimated misspecification is largest. The BC and EB standard errors are both substantially larger than those found by TSLS because they account for specification error in addition to sampling variability. The increases are especially stark in the specifications omitting age controls, which appear to be most heavily misspecified.

The sizable $\|\hat{\Lambda}\hat{\Sigma}\hat{M}'\|$ norms indicate scope for EB to improve upon BC via shrinkage of the individual moment errors. Notably, even in the specifications where $\hat{\sigma}^2 = 0$, EB yields adjustments to BC because the TSLS weighting matrix is potentially suboptimal under heteroscedasticity. However, in practice, the EB standard errors are nearly identical to those of BC, suggesting heteroscedasticity is not a first order concern in these data.

On net, the BC and EB corrections are consistent with season of birth having direct effects on earnings whose sign depends on whether age controls are included: the direct effects are negative when age controls are omitted but positive when age controls are included. The corrections bring TSLS estimates into closer agreement across specifications, yielding returns to schooling in the range of 6--11% per year. This relative stability of the corrected estimates is reassuring, suggesting that the misspecification-aware methods can detect and mitigate the influence of omitted variables bias. Likewise, the elevated standard errors provided by the BC and EB estimators provide a more honest assessment of the composite uncertainty faced by researchers.

9 Conclusion

Econometric models are almost always wrong. As a result, when sample sizes are large, J-tests will tend to reject heavily overidentified models. The methods developed here are designed to address this dilemma by leveraging overidentifying restrictions to correct GMM estimates of parameters of interest.

As the simulations and empirical application illustrate, the proposed correction methods are not entirely automatic. The researcher must have some sense of which aspect of the moment conditions is misspecified. Moreover, the moment conditions must share a common structure that makes exchangeability plausible. As the examples in Section 2 illustrate, the choice of moment scaling determines the exchangeability model and is therefore a substantive modeling decision; alternative scalings may yield different corrections. Finally, the methods developed here are likely to be most useful when realized J-statistics substantially exceed their expected value, as otherwise estimates of σ^2 will be trivially small and $|\mu|$ is likely to be modest.

When all of these conditions are met, the empirical Bayes methods proposed here offer a useful supplement to GMM estimates. Adoption of these methods may carry the additional benefit of improving research transparency by reducing selective reporting of J statistics and other model diagnostics. Rather than fretting about statistical model rejections, researchers should focus their efforts on repairing their estimates of flawed, but ultimately useful, econometric models.

References

Joseph G Altonji and Lewis M Segal. Small-sample bias in gmm estimation of covariance structures. *Journal of Business & Economic Statistics*, 14(3):353–366, 1996.

Isaiah Andrews, Matthew Gentzkow, and Jesse M Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4):1553–1592, 2017.

Isaiah Andrews, Harvey Barnhard, and Jacob Carlson. True and pseudo-true parameters. Technical report, working paper, 2024.

Isaiah Andrews, Jiafeng Chen, and Otavio Tecchio. The purpose of an estimator is what it does: Misspecification, estimands, and over-identification. *arXiv preprint arXiv:2508.13076*, 2025.

Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? *The quarterly journal of economics*, 106(4):979–1014, 1991.

John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

Lawrence D Brown. An ancillarity paradox which appears in multiple linear regression. *The Annals of Statistics*, pages 471–493, 1990.

Kasey S Buckles and Daniel M Hungerman. Season of birth and later outcomes: Old questions, new answers. *Review of economics and statistics*, 95(3):711–724, 2013.

Jiafeng Chen. Empirical bayes when estimation precision predicts parameters. *arXiv preprint arXiv:2212.14444*, 2022.

Victor Chernozhukov, Christian B Hansen, Lingwei Kong, and Weining Wang. Plausible gmm: a quasi-bayesian approach. *arXiv preprint arXiv:2507.00555*, 2025.

Eric Chyn, Brigham Frandsen, and Emily Leslie. Examiner and judge designs in economics: A practitioner's guide. *Journal of Economic Literature*, 63(2):401–439, 2025.

Stefano DellaVigna, John A List, and Ulrike Malmendier. Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, 127(1):1–56, 2012.

Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, 8(4):745–764, 1980.

Pierre-Olivier Gourinchas and Jonathan A Parker. Consumption over the life cycle. *Econometrica*, 70(1): 47–89, 2002.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982.

Lars Peter Hansen and Thomas J. Sargent. *Uncertainty Within Economic Models*, volume 6 of *World Scientific Series in Economic Theory*. World Scientific Publishing Co. Pte. Ltd., Singapore, 2014. ISBN 9789814578110. URL `https://www.worldscientific.com/worldscibooks/10.1142/9028`.

Tjalling C Koopmans and Olav Reiersol. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

Alan B Krueger, Alexandre Mas, and Xiaotong Niu. The evolution of rotation group bias: Will the real unemployment rate please stand up? *Review of Economics and Statistics*, 99(2):258–264, 2017.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Whitney K Newey and Richard J Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Mark R Rosenzweig and Kenneth I Wolpin. Natural "natural experiments" in economics. *Journal of Economic Literature*, 38(4):827–874, 2000.

John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the econometric society*, pages 393–415, 1958.

Appendix

Proof of Lemma 1

Proof. Consider the joint asymptotics of Assumption 2, where $m = m(n) \to \infty$ with $m^2/n \to 0$. By Assumption 2(iii), $\|\hat{M} - M\|_2 = O_p(\sqrt{m/n})$. Estimation error can be decomposed as:

$$\hat{\mu} - \mu = \underbrace{\frac{(M1_m)'Mr}{\|M1_m\|^2} - \mu}_{\text{oracle error}} + \underbrace{\frac{(\hat{M}1_m)'\hat{r}}{\left\|\hat{M}1_m\right\|^2} - \frac{(M1_m)'Mr}{\|M1_m\|^2}}_{\text{first-stage error}}.$$

The oracle error is analyzed as $m \to \infty$ holding population quantities fixed. First, I show the denominator $\|M1_m\|^2$ grows proportionally to m. Write $M = I + G\Lambda$ where $\Lambda = -(G'WG)^{-1}G'W$, so that $M1_m = 1_m + G\Lambda 1_m$. By Assumption 2(iv), $\Lambda 1_m = -(m^{-1}G'WG)^{-1}(m^{-1}G'W1_m) \to 0$ since $(m^{-1}G'WG)^{-1}$ remains bounded and $m^{-1}G'W1_m \to 0$. Each entry of the residual satisfies $|(G\Lambda 1_m)_j| \leq pK \|\Lambda 1_m\| \to 0$ uniformly in j by the uniform bound $|G_{jk}| \leq K$ of Assumption 2(iv). Therefore $(M1_m)_j = 1 + o(1)$ uniformly and $\|M1_m\|^2 = \sum_j (1 + o(1))^2 = m(1 + o(1))$. Moreover, the normalization $\|W\|_2 = 1$ (Assumption 2(ii)) gives $\|W\|_2 = O(1)$. Combined with $\|G\|_2 = O(\sqrt{m})$ from Assumption 2(iv), this gives $\|\Lambda\|_2 = \|(G'WG)^{-1}G'W\|_2 \leq O(1/m) \cdot O(\sqrt{m}) = O(1/\sqrt{m})$ and hence $\|M\|_2 = \|I + G\Lambda\|_2 \leq 1 + O(\sqrt{m}) \cdot O(1/\sqrt{m}) = O(1)$. Next, I evaluate the numerator. Decompose the true local misspecification as $b = \mu 1_m + \sigma\xi$, where μ is the conditional mean bias and ξ is the standardized exchangeable error with $\text{Var}[\xi] = I_m$. Substituting $Mr = M(\mu 1_m + \sigma\xi) + M\varepsilon$ into the numerator yields:

$$(M1_m)'(Mr) = \mu \|M1_m\|^2 + \sigma (M'M1_m)'\xi + (M'M1_m)'\varepsilon,$$

where the expansion uses $(M1_m)'M = (1_m'M')M = 1_m'M'M$ to distribute the inner product. Dividing by the denominator:

$$\mu_{\text{oracle}} - \mu = \sigma \frac{(M'M1_m)'\xi}{\|M1_m\|^2} + \frac{(M'M1_m)'\varepsilon}{\|M1_m\|^2}.$$

Now analyze the two error components on the right-hand side. Because the centered error $\sigma\xi$ has variance $\text{Var}[\sigma\xi] = \sigma^2 I_m$, the variance of the specification error term is $\text{Var}(\sigma(M'M1_m)'\xi) = \sigma^2\|M'M1_m\|^2$. By sub-multiplicativity, $\|M'M1_m\|^2 \leq \|M\|_2^2\|M1_m\|^2 = O(1)\cdot O(m) = O(m)$, using $\|M\|_2 = O(1)$ from above and $\|M1_m\|^2 = O(m)$ from Assumption 2(iv). Chebyshev's inequality then implies the term scales as $O_p(\sqrt{m})$. Dividing by the denominator, the specification error term $\sigma\frac{(M'M1_m)'\xi}{\|M1_m\|^2}$ scales as $O_p(\sqrt{m})/O(m) = O_p(1/\sqrt{m})$. The stochastic term is a linear combination of mean-zero normal variables. Its variance is:

$$\text{Var}\left(\frac{(M'M1_m)'\,\varepsilon}{\|M1_m\|^2}\right) = \frac{(M'M1_m)'\,V\,(M'M1_m)}{\|M1_m\|^4}.$$

The numerator satisfies $(M'M1_m)'V(M'M1_m) \leq \|V\|_2\,\|M\|_2^2\,\|M1_m\|^2 = O(m)$, using $\|V\|_2 = O(1)$ from Assumption (ii) and $\|M\|_2 = O(1)$ established above. Dividing by the squared denominator $\|M1_m\|^4 = O(m^2)$ yields a variance of $O(1/m)$. Consequently, the stochastic error is $O_p(1/\sqrt{m})$. Thus the total oracle error is $O_p(1/\sqrt{m})$. For the first-stage error, the perturbation from replacing M with $\hat{M}$ and r with $\hat{r} = \hat{M}r + o_p(1)$ introduces an additional error that depends on $\|\hat{M} - M\|_2 = O_p(\sqrt{m/n})$. Because the ratio $\hat{\mu}$ is a smooth function of the matrix entries and the denominator $\left\|\hat{M}1_m\right\|^2$ is bounded above and below by positive multiples of m with probability approaching one, standard results on two-step estimation (see, e.g., Section 6 ofNewey and McFadden 1994) yield a first-stage contribution of $O_p(\sqrt{m/n})$. Under $m^2/n \to 0$, this is $o_p(1)$. Combining the two error channels:

$$\hat{\mu} - \mu = O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right).$$

Under $m^2/n \to 0$, both terms vanish and $\sqrt{m/n} = o(1/\sqrt{m})$, so the oracle term $O_p(1/\sqrt{m})$ dominates the first-stage term. □

Proof of Lemma 2

Proof. Consider the joint asymptotics of Assumption 2, where $m = m(n) \to \infty$ with $m^2/n \to 0$. As in Lemma 1, I decompose the estimation error into an oracle component (population M, V with random b, ε) and a first-stage error from replacing population quantities with estimates. The oracle component is the random variable:

$$\widetilde{\sigma^2_{\text{oracle}}} = \frac{\|Mr - \mu_{\text{oracle}}M1_m\|^2 - \text{tr}\,(MVM')}{\text{tr}\,(MM')},$$

where $\mu_{\text{oracle}} = \frac{(M1_m)'Mr}{\|M1_m\|^2}$ and $r = b + \varepsilon$. Write $v = \sigma\xi + \varepsilon$ where $\sigma\xi = b - \mu 1_m$ is the centered specification error and ξ is the standardized exchangeable error, so v is mean-zero with variance $\Sigma = V + \sigma^2 I_m$. I evaluate

the behavior of the oracle estimator as $m \to \infty$. First, expand the residual inside the numerator:

$$Mr - \mu_{\text{oracle}} M1_m = M(\mu 1_m + v) - \mu_{\text{oracle}} M1_m = Mv - (\mu_{\text{oracle}} - \mu) M1_m.$$

Squaring this residual yields the quadratic form:

$$\|Mr - \mu_{\text{oracle}} M1_m\|^2 = v'M'Mv - 2(\mu_{\text{oracle}} - \mu) 1_m' M'Mv + (\mu_{\text{oracle}} - \mu)^2 \|M1_m\|^2.$$

Now bound the terms involving $\mu_{\text{oracle}} - \mu$. By Assumption 2(iv), $\|M1_m\|^2 = O(m)$. From Lemma 1, $\mu_{\text{oracle}} - \mu = O_p(1/\sqrt{m})$. Thus, the third term scales as $O_p(1/m) \times O(m) = O_p(1)$. Because v has bounded variance, $1_m' M'Mv = O_p(\sqrt{m})$, meaning the cross-term scales as $O_p(1/\sqrt{m}) \times O_p(\sqrt{m}) = O_p(1)$. The numerator is therefore dominated by the leading quadratic form $v'M'Mv = \sigma^2 \xi' M'M\xi + 2\sigma \xi' M'M\varepsilon + \varepsilon' M'M\varepsilon$. In the proof of Lemma 1, Assumptions 2(ii) and (iv) were shown to imply $\|M\|_2 = O(1)$, so $\|M'M\|_2 \leq \|M\|_2^2 = O(1)$. Assumption 2(vii) therefore applies directly to the quadratic form $\xi' M'M\xi$, giving $\xi' M'M\xi = \text{tr}(M'M) + O_p(\sqrt{\text{tr}((M'M)^2)})$. Since $\text{tr}((M'M)^2) \leq \|M'M\|_2 \cdot \text{tr}(M'M) = O(1) \cdot O(m) = O(m)$, the remainder is $O_p(\sqrt{m})$, and hence $\sigma^2 \xi' M'M\xi = \sigma^2 \text{tr}(M'M) + O_p(\sqrt{m})$. Since $\varepsilon \sim N(0, V)$, the variance of its quadratic form is $\text{Var}(\varepsilon' M'M\varepsilon) = 2\text{tr}((M'MV)^2)$. By Assumption (ii), $\|V\|_2 = O(1)$, so $\text{tr}((M'MV)^2) \leq \|M'M\|_2^2 \cdot \text{tr}(V^2) \leq \|M'M\|_2^2 \cdot \|V\|_2 \cdot \text{tr}(V) = O(1) \cdot O(1) \cdot O(m) = O(m)$. Chebyshev's inequality therefore yields $\varepsilon' M'M\varepsilon = \text{tr}(MVM') + O_p(\sqrt{m})$. For the cross-term $2\sigma \xi' M'M\varepsilon$, by the independence of ξ and ε, its distribution conditional on ξ is normal with mean zero and variance $4\sigma^2 \xi' M'MVM'M\xi$. Because $\|M'MVM'M\|_2 \leq \|M'M\|_2^2 \cdot \|V\|_2 = O(1)$, Assumption 2(vii) applies directly to $\xi' M'MVM'M\xi$, giving $\xi' M'MVM'M\xi = \text{tr}(M'MVM'M) + O_p(\sqrt{m}) = O_p(m)$ since $\text{tr}(M'MVM'M) \leq \|M'M\|_2^2 \cdot \text{tr}(V) = O(m)$. The conditional variance is therefore $O_p(m)$, meaning the cross-term $\sigma \xi' M'M\varepsilon$ itself evaluates to $O_p(\sqrt{m})$. Therefore, the combined quadratic form evaluates to:

$$v'M'Mv = \sigma^2 \text{tr}(MM') + \text{tr}(MVM') + O_p(\sqrt{m}).$$

Substitute this back into the expression for the asymptotic variable $\widetilde{\sigma^2_{\text{oracle}}}$:

$$\widetilde{\sigma^2_{\text{oracle}}} = \frac{\sigma^2 \text{tr}(MM') + \text{tr}(MVM') + O_p(\sqrt{m}) + O_p(1) - \text{tr}(MVM')}{\text{tr}(MM')}.$$

The $\text{tr}(MVM')$ terms cancel perfectly. I previously established that the denominator satisfies $\text{tr}(MM') \geq \frac{(m-p)^2}{m}$. Because the parameter dimension p is fixed, the denominator is strictly bounded below by a term

of order $O(m)$. Dividing the $O_p(\sqrt{m})$ remainder by a denominator that grows at the rate m gives:

$$\widetilde{\sigma^2_{\text{oracle}}} = \sigma^2 + O_p\left(\frac{1}{\sqrt{m}}\right).$$

For the first-stage error from replacing $(M, V, \mu_{\text{oracle}})$ with $(\hat{M}, \hat{V}, \hat{\mu})$, Assumption 2(iii) provides that $\|\hat{M} - M\|_2 = O_p(\sqrt{m/n})$ and $\|\hat{V} - V\|_2 = O_p(\sqrt{m/n})$. Because the estimator is a smooth function of matrix traces and the denominator $\text{tr}(\hat{M}\hat{M}')$ is bounded above and below by positive multiples of m with probability approaching one, the first-stage perturbation contributes $O_p(\sqrt{m/n})$. Since the true parameter $\sigma^2 \geq 0$, truncation at zero preserves the rate. Combining yields the result:

$$\widehat{\sigma^2} - \sigma^2 = O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right).$$

□

Proof of Lemma 3

Proof. **Part 1: Risk improvement iff $M\Sigma\Lambda' \neq 0$.**

The error of θ^*_{BC} is $\sqrt{n}(\theta^*_{BC} - \theta_0) = \Lambda(\sigma\xi + \varepsilon)$, giving $R(\theta^*_{BC}) = \Lambda\Sigma\Lambda'/n$. The error of θ^* is $\sqrt{n}(\theta^* - \theta_0) = \Lambda(I - \Pi)(\sigma\xi + \varepsilon)$ where $\Pi = \Sigma M'(M\Sigma M')^+ M$. Since Π is the BLP coefficient for r given Mr, the residual variance is $(I - \Pi)\Sigma(I - \Pi)' = \Sigma - \Sigma M'(M\Sigma M')^+ M\Sigma$, giving $R(\theta^*) = \Lambda\left(\Sigma - \Sigma M'(M\Sigma M')^+ M\Sigma\right)\Lambda'/n$. Therefore:

$$R(\theta^*_{BC}) - R(\theta^*) = \frac{1}{n}\Lambda\Sigma M'\left(M\Sigma M'\right)^+ M\Sigma\Lambda' + o(1/n).$$

This matrix is manifestly positive semi-definite, as it equals $(1/n) \cdot A'(M\Sigma M')^+ A$ where $A = M\Sigma\Lambda'$ and $(M\Sigma M')^+$ is positive semi-definite. Since Assumption 2(vi) ensures $(M\Sigma M')^+$ is positive definite on the range of M, this $p \times p$ matrix is positive semidefinite and nonzero if and only if $M\Sigma\Lambda' \neq 0$.

More generally, $\Lambda\Sigma M'(M\Sigma M')^+ M\Sigma\Lambda'$ is positive definite if and only if for all $x \neq 0$, $(M\Sigma\Lambda' x)'(M\Sigma M')^+(M\Sigma\Lambda' x) > 0$. Since $(M\Sigma M')^+$ is positive definite on range(M) and $M\Sigma\Lambda' x \in$ range(M), this holds if and only if $M\Sigma\Lambda' x \neq 0$ for all $x \neq 0$, i.e., rank$(M\Sigma\Lambda') = p$. Since $M\Sigma\Lambda'$ is $m \times p$ with rank at most $\min(m - p, p)$, positive definiteness requires $m - p \geq p$, hence $m \geq 2p$. When $p = 1$, the nonzero and positive definite conditions coincide.

Part 2: Equivalent characterizations of $M\Sigma\Lambda' = 0$.

Using $\Lambda' = -WG(G'WG)^{-1}$:

$$M\Sigma\Lambda' = -M\Sigma WG(G'WG)^{-1}.$$

Because $(G'WG)^{-1}$ is invertible, $M\Sigma\Lambda' = 0$ if and only if $M\Sigma WG = 0$, i.e., $\Sigma WG \in \text{col}(G)$.

Item i): $p = 1$.

Since $p = 1$, G is an $m \times 1$ vector. Then $M\Sigma WG = \Sigma WG - G \cdot \frac{G'W\Sigma WG}{G'WG}$. This vanishes if and only if $\Sigma WG \propto G$.

Item ii): General p.

The condition $M\Sigma WG = 0$ holds if and only if $\Sigma WG = G \underbrace{(G'WG)^{-1}G'W\Sigma WG}_{=: C}$, i.e., there exists a $p \times p$ matrix C such that $\Sigma WG = GC$. □

Proof of Proposition 2

Proof. Consider the joint asymptotics of Assumption 2, where $m = m(n) \to \infty$ with $m^2/n \to 0$. The prediction error decomposes into a hyperparameter channel, a first-stage matrix estimation channel, and a sensitivity estimation error channel.

Hyperparameter channel.

At the population matrices (M, Σ), the scalar errors $|\hat{\mu} - \mu| = O_p(1/\sqrt{m})$ and $|\widehat{\sigma^2} - \sigma^2| = O_p(1/\sqrt{m})$ from Lemmas 1-- 2 propagate through the linear projection. Since the projection $\Sigma M'(M\Sigma M')^+(Mr - \mu M 1_m)$ has norm bounded by $O_p(\sqrt{m})$, the total vector error from hyperparameter estimation is $\|\hat{L} - L^*\|_2 = O_p(1)$, giving an average element-wise rate of $O_p(1/\sqrt{m})$.

Matrix estimation channel.

The first-stage errors $\|\hat{M} - M\|_2 = O_p(\sqrt{m/n})$, $\|\hat{V} - V\|_2 = O_p(\sqrt{m/n})$ perturb the Moore--Penrose pseudoinverse $(M\Sigma M')^+$. (The $\hat{\sigma}^2$ component of $\hat{\Sigma}$ converges at the slower rate $O_p(1/\sqrt{m})$ and is handled in the hyperparameter channel above.) By Assumption 2(vi), the nonzero eigenvalues of $M\Sigma M'$ are bounded away from zero, so standard two-step estimation results yield $\|(\hat{M}\hat{\Sigma}\hat{M}')^+ - (M\Sigma M')^+\|_2 = O_p(\sqrt{m/n})$. Because the residual vector $Mr - \mu M 1_m$ has norm $O_p(\sqrt{m})$ (using $\|M\|_2 = O(1)$) and $\|\Sigma M'\|_2 = O(1)$, propagating through the prediction formula gives a vector perturbation of norm $O_p(\sqrt{m} \cdot \sqrt{m/n}) = O_p(m/\sqrt{n})$. Dividing by $\sqrt{m}$ yields an element-wise contribution of $O_p(\sqrt{m/n})$.

Downstream EB error.

Since both $\hat{\theta}_{EB}$ and θ^* use the sample sensitivity matrix $\hat{\Lambda}$,

$$\hat{\theta}_{EB} - \theta^* = -\frac{1}{\sqrt{n}}\hat{\Lambda}(\hat{L} - L^*).$$

Now $\hat{\Lambda} = \Lambda + O_p(1/\sqrt{n})$ by Assumption 2(iii), and $\|\Lambda\|_2 = O(1/\sqrt{m})$ since $\|\Lambda\|_2 \le \|(G'WG)^{-1}\|_2 \, \|G'W\|_2 = O(1/m) \cdot O(\sqrt{m})$ by Assumption 2(iv) and the normalization $\|W\|_2 = 1$. Hence $\|\hat{\Lambda}\|_2 = O_p(1/\sqrt{m})$. The

vector prediction error $\hat{L} - L^*$ has norm $O_p(1) + O_p(m/\sqrt{n})$ from the hyperparameter and matrix estimation channels above. Multiplying:

$$\hat{\theta}_{EB} - \theta^* = O_p\left(\frac{1}{\sqrt{mn}}\right) + O_p\left(\frac{\sqrt{m}}{n}\right).$$

□

Proof of Theorem 1

Proof. I work conditional on (υ, ς) under Assumptions 1--3.

Leading-term reduction.

From the proof of Proposition 1, $\sqrt{n}(\hat{\theta}_{BC} - \theta_0) = \Lambda\eta - (\hat{\mu} - \mu)\Lambda 1_m + R_{BC}$ where R_{BC} collects remainders of order $O_p(\sqrt{m/n})$ from first-stage estimation. The oracle mean estimator satisfies $\hat{\mu}_{\text{oracle}} - \mu = w'\eta$, and the feasibility gap $|\hat{\mu} - \hat{\mu}_{\text{oracle}}| = O_p(\sqrt{m/n})$ (the first-stage error from Lemma 1). Setting $B = \Lambda(I_m - 1_m w')$ yields

$$\sqrt{n}(\hat{\theta}_{BC} - \theta_0) = B\eta + \tilde{R}_{BC}.$$

The remainder $\tilde{R}_{BC}$ collects: (a) the sensitivity estimation error $(\hat{\Lambda} - \Lambda)\eta = O_p(\sqrt{m}/n^{1/2})$, which upon premultiplication by $V_{BC}^{-1/2}$ (with $\|V_{BC}^{-1/2}\|_{\text{op}} = O(\sqrt{m})$) becomes $O_p(m/\sqrt{n}) = o_p(1)$ (since $m^2/n \to 0$ by Assumption 2(i)); and (b) the mean estimation feasibility gap $(\hat{\mu} - \hat{\mu}_{\text{oracle}})\Lambda 1_m$, of order $O_p(\sqrt{m/n})$ (since $\|\Lambda 1_m\| = O(1)$), which upon standardization by $V_{BC}^{-1/2}$ becomes $O_p(m/\sqrt{n}) = o_p(1)$. Hence $V_{BC}^{-1/2}\tilde{R}_{BC} = o_p(1)$.

Exchangeable CLT.

Write $\eta = \sigma\xi + \varepsilon$ with $\xi \perp \varepsilon$. Since $\varepsilon \sim N(0, V)$, $B\varepsilon \sim N(0, BVB')$. It suffices to show $\sigma B\xi \to_d N(0, \sigma^2 BB')$. Denoting the j-th column of B by B_j, write $\sigma B\xi = \sigma \sum_{j=1}^m B_j \xi_j$. Since $B1_m = 0$, the columns sum to zero. By the Cramér–Wold device, fix any unit vector $c \in \mathbb{R}^p$ and consider $T_m = \sigma \sum_j (c'B_j)\xi_j$. This is a sum of m centered, uncorrelated random variables with $\text{Var}(T_m) = \sigma^2 \sum_j (c'B_j)^2 = \sigma^2 c'BB'c$. Write $a_j = c'B_j$. Since $B1_m = 0$, the weights satisfy $\sum_j a_j = 0$. Assumption 2(iv) ensures $\max_j a_j^2 / \sum_k a_k^2 \to 0$. To see this, the j-th column of $B = \Lambda(I_m - 1_m w')$ is $B_j = \Lambda(e_j - w_j 1_m)$, so $a_j = \lambda_j - w_j\delta_m$ where $\lambda_j = (c'\Lambda)_j$ and $\delta_m = c'\Lambda 1_m = o(1)$ by Assumption 2(iv). Since $\Lambda = -(G'WG)^{-1}G'W$, write $b = c'(m^{-1}G'WG)^{-1}$ and $\lambda_j = -m^{-1}(WGb')_j$. The vector b is $O(1)$ by the eigenvalue bounds in Assumption 2(iv). Writing $v = Gb'$, the bounded-entry condition gives $|v_j| \le \sqrt{p}\, K\|b\| = O(1)$. For the numerator, the bounded row sums of W in Assumption 2(ii) give $\max_j |\lambda_j| = m^{-1} \max_j |\sum_k W_{jk} v_k| \le m^{-1} (\max_j \sum_k |W_{jk}|) (\max_k |v_k|) = O(1/m)$. For the denominator, Cauchy--Schwarz gives $\sum_j \lambda_j^2 = m^{-2}\|Wv\|^2 \ge m^{-2}(v'Wv)^2/\|v\|^2$. Now $v'Wv = mb(m^{-1}G'WG)b' \ge m\Delta^{-1}\|b\|^2$ by Assumption 2(iv), and $\|v\|^2 \le mpK^2\|b\|^2$, so $\sum_j \lambda_j^2 \ge \delta^2/(pK^2\Delta^2 m)$.

Hence $\max_j \lambda_j^2 / \sum_k \lambda_k^2 = O(1/m) \to 0$. Since $w_j = O(1/m)$ and $\delta_m = o(1)$, the correction $w_j \delta_m = o(1/m)$ is negligible relative to the λ_j of typical magnitude $O(1/\sqrt{m})$, and the Lindeberg condition transfers to $\{a_j\}$. I verify the moment conditions for the method of moments directly, without appeal to independence.

By exchangeability, $E[\xi_i^2 \xi_j^2]$ is constant across all pairs $i \neq j$; denote this common value by ϱ_m and note that it may be a function of m. Applying Assumption 2(vii) with $A = I_m/m$ gives $\mathrm{Var}(\|\xi\|^2/m) \leq C'/m$, since $E[\xi' A \xi] = \mathrm{tr}(A) = 1$ by the unit-variance normalization and $\mathrm{tr}(A^2) = 1/m$. Expanding the left-hand side by exchangeability:

$$\mathrm{Var}(\|\xi\|^2/m) = O(1/m) + \left(1 - \frac{1}{m}\right) \mathrm{Cov}(\xi_1^2, \xi_2^2),$$

where the $O(1/m)$ term collects the diagonal contributions $m^{-2} \sum_j \mathrm{Var}(\xi_j^2) \leq m^{-1} \sup_j E[\xi_j^4]$, which is bounded by Assumption 3. Combining with the upper bound from (vii) yields $\mathrm{Cov}(\xi_1^2, \xi_2^2) = O(1/m)$. Since $E[\xi_j^2] = 1$ for all j,

$$\varrho_m = E[\xi_1^2 \xi_2^2] = \mathrm{Cov}(\xi_1^2, \xi_2^2) + 1 = 1 + O(1/m).$$

For the fourth moment of T_m, I expand $E[(a'\xi)^4]$ by exchangeability and bound the cross-moments in which any index appears to an odd power. By the quantitative finite de Finetti theorem (Diaconis and Freedman 1980), the 4-marginal of an m-exchangeable vector is within $O(1/m)$ total variation of a mixture of i.i.d. measures with mixing measure ν. Under any i.i.d. component with mean μ_θ, $E_\theta[\xi_i \xi_j] = \mu_\theta^2$ for $i \neq j$. Since the TV bound and the uniform moment bounds of Assumption 3 – the latter controlling the TV error for unbounded integrands via uniform integrability – give $|E[\xi_i \xi_j] - \int \mu_\theta^2 \, d\nu| = o(1)$, and $E[\xi_i \xi_j] = 0$ by construction, it follows that $\int \mu_\theta^2 \, d\nu = o(1)$. Consequently, for distinct indices i, j, k, $E_\nu[\xi_i^2 \xi_j \xi_k] = \int E_\theta[\xi_i^2] \, \mu_\theta^2 \, d\nu \leq \left(\int E_\theta[\xi_i^4] \, d\nu\right)^{1/2} \left(\int \mu_\theta^4 \, d\nu\right)^{1/2} = O(1) \cdot o(1) = o(1)$, and similarly $E_\nu[\xi_i \xi_j \xi_k \xi_l] = \int \mu_\theta^4 \, d\nu = o(1)$. To see this, split at $|\mu_\theta| = 1$: on $\{|\mu_\theta| \leq 1\}$, $\mu_\theta^4 \leq \mu_\theta^2$ gives $\int \leq \int \mu_\theta^2 \, d\nu = o(1)$; on $\{|\mu_\theta| > 1\}$, Jensen gives $\mu_\theta^4 \leq E_\theta[\xi_i^4]$ and the integral is $o(1)$ by uniform integrability: the uniform moment bounds of Assumption 3 give $\sup_m \int E_\theta[\xi_i^4] \, d\nu_m < \infty$, and $\nu_m(\{|\mu_\theta| > 1\}) \to 0$. Adding the TV error: $|E[\xi_i^3 \xi_j]|$, $|E[\xi_i^2 \xi_j \xi_k]|$, $|E[\xi_i \xi_j \xi_k \xi_l]| = o(1)$ for distinct indices. Writing $s^2 = \mathrm{Var}(T_m)$, their contribution to $E[T_m^4]/s^4$ is $o(1)$: the zero-sum constraint converts the type-[3,1] coefficient sum to $-\sum a_j^4$, which is $o(s^4)$ by the Lindeberg condition (which follows from Assumption 2(iv)), while the type-[2,1,1] and type-[1,1,1,1] sums each contribute $o(1)$, since the zero-sum constraint collapses their coefficient sums to $o(s^4)$ by the same Lindeberg argument. The dominant terms give:

$$E[T_m^4]/(\mathrm{Var}\, T_m)^2 = 3\varrho_m + (\kappa_4 + 3 - 3\varrho_m) \cdot \frac{\sum_j a_j^4}{(\sum_j a_j^2)^2}$$

where $\kappa_4 = E[\xi_j^4] - 3$. Since $\varrho_m \to 1$ and $\sum a_j^4/(\sum a_j^2)^2 \to 0$ by the Lindeberg condition, the ratio converges to 3.

The same argument extends to all even moments. At each order $2k$, expanding $E[(T_m/s)^{2k}]$ by exchangeability yields a sum over partition types of $\{1, \ldots, 2k\}$ into groups assigned to distinct indices. The dominant type is the fully-paired partition $[2, \ldots, 2]$ (k pairs), which contributes $(2k-1)!!$ terms with moment factor $\varrho_m^{(k)} := E[\xi_{i_1}^2 \cdots \xi_{i_k}^2]$ for k distinct indices. To verify $\varrho_m^{(k)} = 1 + o(1)$, expand $E[(\|\xi\|^2/m)^k] = m^{-k} \sum_{i_1,\ldots,i_k} E[\xi_{i_1}^2 \cdots \xi_{i_k}^2]$. By exchangeability, the $(m)_k$ tuples with all k indices distinct each contribute $\varrho_m^{(k)}$, while the remaining $O(m^{k-1})$ tuples with at least one repeated index each contribute at most $\sup_m E[|\xi_j|^{2k}] < \infty$ by Assumption 3. Hence $E[(\|\xi\|^2/m)^k] = ((m)_k/m^k)\, \varrho_m^{(k)} + O(1/m)$. Meanwhile, $\|\xi\|^2/m \to_p 1$ by Assumption 2(vii), and $\sup_m E[(\|\xi\|^2/m)^k] < \infty$ by the uniform moment bounds, so $E[(\|\xi\|^2/m)^k] \to 1$. Since $(m)_k/m^k = 1 - O(k^2/m)$, solving gives $\varrho_m^{(k)} = 1 + o(1)$ for each fixed k. All partition types with at least one index appearing to an odd power contribute $o(1)$ by the Diaconis–Freedman and Lindeberg arguments. For partitions where each odd index appears exactly once (e.g., type [3,1]), the cross-moments are $o(1)$ by the Cauchy–Schwarz and uniform integrability arguments above. For partitions with higher odd powers (e.g., type [3,3]), the cross-moments may be $O(1)$, but the associated coefficient sums are $o(s^{2k})$ by the zero-sum and Lindeberg conditions. Hence $E[(T_m/\sqrt{\operatorname{Var} T_m})^{2k}] \to (2k-1)!!$ for all k. By the method of moments (which requires no independence), $T_m/\sqrt{\operatorname{Var}(T_m)} \to_d N(0,1)$.

Combining.

By independence of ξ and ε,

$$B\eta = \sigma B\xi + B\varepsilon \to_d N(0, \sigma^2 BB' + BVB') = N(0, B\Sigma B') = N(0, V_{BC}).$$

After standardization, $V_{BC}^{-1/2} B\eta \to_d N(0, I_p)$. Since $V_{BC}^{-1/2} \tilde{R}_{BC} = o_p(1)$, Slutsky's theorem gives the result.

□

Proof of Corollary 1

Proof. The proof parallels that of Theorem 1, with additional verification that the $\hat{\sigma}^2$ correction is asymptotically negligible.

Leading-term reduction.

Define $A = \Lambda(I_m - \Pi)$ where $\Pi = \Sigma M'(M\Sigma M')^+ M$. The oracle EB estimator satisfies $\sqrt{n}(\theta^* - \theta_0) = A\eta$. The $\hat{\mu}$ correction contributes $-(w'\eta)A1_m$, giving $B_{EB} = A(I_m - 1_m w')$. The $\hat{\sigma}^2$ correction arises through the dependence of $\hat{\Pi}$ on $\hat{\sigma}^2$: the leading-order effect on $\sqrt{n}(\hat{\theta}_{EB} - \theta_0)$ is of order $|\hat{\sigma}^2 - \sigma^2| \cdot \|\Lambda \frac{\partial \Pi}{\partial \sigma^2} \eta\|$.

Since $\frac{\partial \Pi}{\partial \sigma^2}$ is bounded and $\operatorname{tr}\left(\Lambda \frac{\partial \Pi}{\partial \sigma^2} \Sigma \left(\frac{\partial \Pi}{\partial \sigma^2}\right)' \Lambda'\right) = O(1/m)$, one has $\|\Lambda \frac{\partial \Pi}{\partial \sigma^2} \eta\| = O_p(1/\sqrt{m})$. Combined with $|\hat{\sigma}^2 - \sigma^2| = O_p(1/\sqrt{m})$, the product is $O_p(1/m) = o_p(1/\sqrt{m})$. Matrix estimation corrections $(\hat{M} - M, \hat{V} - V)$ contribute $O_p(\sqrt{m/n})$ as in Proposition 2. Hence

$$\sqrt{n}(\hat{\theta}_{EB} - \theta_0) = B_{EB}\eta + o_p(1/\sqrt{m}).$$

Exchangeable CLT.

The exchangeable CLT for $B_{EB}\xi$ follows identically to the argument for Theorem 1, since B_{EB} shares the column-sum-zero property: $B_{EB}1_m = A(I_m - 1_m w')1_m = A \cdot 0 = 0$, since $w'1_m = 1$. Assumption 2(iv) again guarantees the Lindeberg condition.

Combining.

Combining via Slutsky as before yields the result. For the asymptotic variance comparison, write $V_{BC} - V_{EB} = B\Sigma B' - B_{EB}\Sigma B'_{EB}$ where $B = \Lambda\mathcal{C}$, $B_{EB} = \Lambda(I_m - \Pi)\mathcal{C}$, and $\mathcal{C} = I_m - 1_m w'$. By definition of $\Sigma = V + \sigma^2 I_m$,$\mathcal{C}\Sigma\mathcal{C}' = \mathcal{C}V\mathcal{C}' + \sigma^2\mathcal{C}\mathcal{C}' =: \Omega$. Consequently,

$$V_{BC} - V_{EB} = \Lambda\left[\Omega - (I_m - \Pi)\Omega(I_m - \Pi)'\right]\Lambda'.$$

For the relationship between feasible and oracle asymptotic variances, write $\mathcal{C} = I_m - 1_m w'$ where w has entries of order $O(1/m)$. Then $V_{BC} = \Lambda\mathcal{C}\Sigma\mathcal{C}'\Lambda' = \Lambda\Sigma\Lambda' + O(1/m)$ and $V_{EB} = \Lambda(I_m - \Pi)\mathcal{C}\Sigma\mathcal{C}'(I_m - \Pi)'\Lambda' = \Lambda(I_m - \Pi)\Sigma(I_m - \Pi)'\Lambda' + O(1/m)$, where the $O(1/m)$ bounds follow from $\|\Lambda\|_2 = O(1/\sqrt{m})$ and the rank-one structure of $\mathcal{C} - I_m = -1_m w'$. By Lemma 3, the oracle variance difference $\Lambda\Sigma\Lambda' - \Lambda(I_m - \Pi)\Sigma(I_m - \Pi)'\Lambda'$ is positive semidefinite and of order $O(1/m)$, so the oracle risk comparison characterizes the leading-term behavior of $V_{BC} - V_{EB}$. □