WORKING PAPERS

N° 1614

September 2026

A Neyman-Orthogonalization Approach to The Incidental Parameter Problem In Likelihood Models

Stéphane Bonhomme, Koen Jochmans and Martin Weidner

Toulouse School of Economics

A NEYMAN-ORTHOGONALIZATION APPROACH TO THE INCIDENTAL PARAMETER PROBLEM IN LIKELIHOOD MODELS*

Stéphane Bonhomme†
Department of Economics, University of Chicago

Koen Jochmans‡
Toulouse School of Economics, Université Toulouse Capitole

Martin Weidner§
Department of Economics and Nuffield College, University of Oxford

September 1, 2026

Abstract

A popular approach to perform inference on a target parameter in the presence of nuisance parameters is to construct estimating equations that are orthogonal to the nuisance parameters, in the sense that their expected first derivative is zero. Such first-order orthogonalization allows the estimator of the nuisance parameters to converge at a slower-than-parametric rate. It may, however, not suffice when the nuisance parameters are very imprecisely estimated. Leading examples are models for panel and network data that feature fixed effects. In this paper, we show how, in the conditional-likelihood setting, estimating equations can be constructed that are orthogonal to any chosen order q, in that their leading q expected derivatives are zero. This yields estimators of target parameters that are unaffected by the presence of nuisance parameters to order q. In an empirical illustration, we apply our method to a fixed-effect model of team production.

JEL Classification: C13, C23, C55.

Keywords: Neyman-orthogonality, incidental parameter, higher-order bias correction, networks.

*We are grateful to the Editor and referees, as well as Dmitry Arkhangelsky, Jin Hahn, Bo Honoré, Roger Moon, Whitney Newey, Andres Santos, Vira Semenova, and Vasilis Syrgkanis for comments and discussion.
Funded by the European Union (ERC-NETWORK-101044319) and by the French Government and the French National Research Agency under the Investissements d'Avenir program (ANR-17-EURE-0010). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

†sbonhomme@uchicago.edu
‡koen.jochmans@tse-fr.eu
§martin.weidner@economics.ox.ac.uk

1 Introduction

Inference in the presence of nuisance parameters has received substantial attention. One fruitful way to proceed is to work with estimating equations that are orthogonal with respect to the nuisance parameters in the sense of Neyman (1959). Such equations underlie many of the results in semiparametric estimation (Newey, 1994) and are at the heart of recent advances on doubly-robust estimation and high-dimensional inference as discussed in, for example, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). A key finding is that Neyman-orthogonality permits the construction of asymptotically unbiased estimators that converge at the usual $n^{-1/2}$-rate provided the nuisance parameter has a convergence rate that is faster than $n^{-1/4}$, where n is the sample size.

However, the faster-than-$n^{-1/4}$ requirement often fails in problems where the dimension of the nuisance parameter is large relative to the sample size. Examples are panel data models with fixed effects, which are widely used in linear and nonlinear difference-in-differences settings. There, we observe N units over T periods of time and the model includes both common parameters and unit-specific nuisance parameters (such as heterogeneous intercepts or slopes). The latter are estimated at the rate $T^{-1/2}$. For an estimator of the former based on Neyman-orthogonalization to be successful we would therefore need that $T^{-1/2} = o((NT)^{-1/4})$, which translates into the requirement that $N = o(T)$. This is usually not a realistic condition in microeconometric applications. In fact, under this requirement the standard fixed-effect estimator would permit asymptotically-valid inference. Consequently, (first-order) Neyman-orthogonalization does not solve the incidental parameter problem in panel data.[1]

The issue can be even more severe in high-dimensional regressions on network data. In such settings, the convergence rate of the estimator of the nuisance parameter depends on the connectivity structure of the network (Jochmans and Weidner, 2019). Examples include the estimation of teacher value-added (Jackson, Rockoff and Staiger, 2014), of the contributions of worker and firm heterogeneity to the variance of log wages and other covariance components (Abowd, Kramarz and Margolis,

[1]The problem is reminiscent of the poor performance of double machine-learning techniques in some settings, as recently documented by Wüthrich and Zhu (2021) and Angrist and Frandsen (2022). A related problem where the conventional approach was formally shown to fail is a nonlinear version of the judge-leniency design, see Hahn and Hausman (2021).

1999, Kline, Saggio and Sølvsten, 2020), as well as of complementarity patterns in team production (Ahmadpoor and Jones, 2019, Bonhomme, 2021). Fixed effects in network-formation models are also poorly estimated, especially in the prevalent case where the network is sparse (see, e.g., Graham, 2020).

Motivated by these concerns, we focus on a higher-order generalization of Neyman-orthogonality that was proposed by Mackey, Syrgkanis and Zadik (2018). An estimating equation is Neyman-orthogonal to order q when all q leading derivatives with respect to the nuisance parameter have zero expectation. When $q = 1$, this means that the expected Jacobian is zero, which recovers the conventional notion of Neyman-orthogonality (to order one). Working with estimating equations that are Neyman-orthogonal to order q, when combined with sample splitting, allows one to construct asymptotically-linear estimators when nuisance parameters are estimated at a rate faster than $n^{-1/2(q+1)}$. As an example, in the panel data problem, this yields valid inference under the requirement that $N = o(T^q)$. Combining orthogonalization with sample splitting (or cross-fitting) is important to achieve such an improvement, because orthogonalized estimating equations, by themselves, do not, in general, deliver estimators with improved sampling properties.

Working in the conditional-likelihood setting, we show how to construct estimating equations that are orthogonal to any chosen order. These estimating equations can be understood to be generalizations of the projected score of Small and McLeish (1989) and Waterman and Lindsay (1996). They can also be seen as higher-order influence functions, as in Robins, Li, Tchetgen Tchetgen and van der Vaart (2008) and van der Vaart (2014). Our approach applies to general low-dimensional target parameters that satisfy a moment restriction. This includes functions of the nuisance parameters such as average elasticities or other average effects. The conditional-likelihood setting is crucial for our approach, since it allows us to orthogonalize a given estimating equation without introducing additional nuisance parameters. Avoiding such additional nuisance parameters is not essential to achieve orthogonality to order one, as is well known. It is, however, what enables the construction of higher-order orthogonalized estimating equations, as we explain in Subsection 4.2. We further discuss the non-likelihood case in Section 8.

Our approach relates to techniques to correct for bias in panel data (Hahn and Newey, 2004, Dhaene and Jochmans, 2015a,b) and to the literature on small measurement error (Chesher, 1991, Evdokimov and Zeleneev, 2023). However, in contrast

with these approaches, we do not restrict the nuisance parameters beyond the fact that they can be estimated at a certain rate.

We illustrate the usefulness of our approach in several examples and in an empirical application to the estimation of nonlinear regressions on network data; a problem for which, at present, no alternative solutions exist. In this setting, we estimate a constant elasticity of substitution (CES) production function from the scientific output of research collaborations. As in Ahmadpoor and Jones (2019), the production function depends on researcher-specific fixed effects. Estimates of the parameters can be used to quantify the degree of complementarity among researchers within teams, and to compute the impact of counterfactual re-allocations in the spirit of earlier work by Graham, Imbens and Ridder (2014).

This problem is difficult because in the data that we use (taken from Ductor, Fafchamps, Goyal and Van der Leij, 2014 and concerning publications in economics on EconLit), the number of collaborations per researcher is quite low. A conventional estimator is thus likely to suffer from substantial bias. Our procedure uncovers the presence of complementarity among authors in the production of research articles. In a counterfactual exercise we also find that randomly pairing researchers would lead to a decrease in the average quality of articles. Our findings are corroborated in a simulation experiment targeted to our empirical application.

2 Problem statement and motivation

2.1 Setup

Let $Z_i = (Y_i, X_i)$ be random vectors, for $i = 1, .., N$. We consider a setting where the conditional density function of Y_i at y given $X_i = x$, say $\ell(y \,|\, x; \theta_0, \eta_{i0})$, is known up to the parameters θ_0 and η_{i0}. Throughout, we will treat $\eta_{10}, \ldots, \eta_{N0}$ as nuisance parameters, and leave the marginal density of the conditioning variable, $\ell_{X_i}(x)$, unrestricted. We are interested in estimating a parameter μ_0 that is defined through the moment condition

$$\mathbb{E}\left(\sum_{i=1}^{N} u(Z_i; \theta_0, \eta_{i0}, \mu_0)\right) = 0, \tag{2.1}$$

where the expectations are over Z_i under $\ell(y \mid x; \theta_0, \eta_{i0})\, \ell_{X_i}(x)$. While the function u could additionally depend on i, for instance in settings where the dimension of η_{i0} differs across i, we omit this dependence for conciseness. We assume that, for all $i = 1, \ldots, N$, Z_i contains n_i individual observations, and denote the total number of observations as $n = \sum_{i=1}^{N} n_i$. For example, in a balanced panel data setting with N units and T time periods, Z_i is the time series of unit i's observations, $n_i = T$ for all i, and $n = NT$.

Our setup accommodates different types of target parameters. As an example, we can set $\mu_0 = \theta_0$. In this case, using $u(z; \theta, \eta_i)$ as a shorthand for $u(z; \theta, \eta_i, \theta)$, one possibility is to use the score,

$$u(z; \theta, \eta_i) = \frac{\partial \log \ell(y \mid x; \theta, \eta_i)}{\partial \theta}.$$

More generally, the moment condition (2.1) defines the target parameter

$$\mu_0 = \mu(\theta_0, \eta_{10}, \ldots, \eta_{N0}, \ell_{X_1}, \ldots, \ell_{X_N}),$$

which can be a function of the parameters θ_0 and η_{i0} describing the conditional distribution of Y_i given X_i, of the marginal distribution of X_i, and (implicitly) of the sample size. For example, we may be interested in an average effect of the form

$$\mu_0 = \frac{1}{N} \sum_{i=1}^{N} \int m(x; \theta_0, \eta_{i0}) \ell_{X_i}(x)\, dx,$$

where m is a known function.

To illustrate the setup we will refer to two leading examples.

Example: Neyman-Scott model. Our first example is the well-known Neyman and Scott (1948) model. Here,

$$Y_{ij} = \eta_{i0} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{iid } \mathcal{N}\left(0, \sigma_0^2\right), \quad i = 1, \ldots, N, \quad j = 1, \ldots, T, \tag{2.2}$$

and the goal is to estimate $\theta_0 = \sigma_0^2$ in the presence of the nuisance parameters $\eta_{10}, \ldots, \eta_{N0}$. Define, for all $i = 1, \ldots, N$,

$$u(Y_i; \sigma^2, \eta_i) = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{T} (Y_{ij} - \eta_i)^2, \tag{2.3}$$

where $Y_i = (Y_{i1}, \ldots, Y_{iT})^\top$ has dimension $n_i = T$, and the total number of observations is $n = NT$. It is well-known that the maximum-likelihood estimator of σ_0^2 is on average too small, suffering from bias $-\sigma_0^2/T$. While in this panel data problem first-order orthogonality does not reduce the order of this bias, we demonstrate below that second-order orthogonalization fully removes it.

Example: CES production function. Consider an environment where we observe workers producing output in n teams of size 2. Moreover, let $k(j, 1)$ and $k(j, 2)$ denote the workers in team j, and write $\mathcal{K} = \{(k(j, 1), k(j, 2)) : j = 1, \ldots, n\}$ for the set of workers in all teams; note that a given worker may be part of multiple teams. Consider a model for team production where team output is a CES aggregate of worker inputs (as in Ahmadpoor and Jones, 2019),

$$Y_j = \left(\frac{\eta_{k(j,1)0}^{\gamma_0} + \eta_{k(j,2)0}^{\gamma_0}}{2} \right)^{\frac{1}{\gamma_0}} \varepsilon_j^{\sigma_0}, \quad \log \varepsilon_j \,|\, \mathcal{K} \sim \text{iid } \mathcal{N}(0, 1), \quad j = 1, \ldots, n. \tag{2.4}$$

In this model, one may be interested in estimating the substitution parameter γ_0 or the log error variance σ_0^2, average elasticities, or effects of counterfactual re-allocations of workers to teams, for example.[2]

To analyze this example we consider $N \leq n$ subsets of teams j, of size n_i each. Let Y_i denote the vector of team outcomes in subset i, let $\mathcal{K}_i$ denote the set of indicators for workers belonging to those teams, and let η_i be the collection of all fixed effects of those workers. Finally, let $\theta = (\gamma, \sigma^2)^\top$. The scores with respect to γ and σ^2 take the form $u(Y_i, \mathcal{K}_i; \theta, \eta_i)$. In contrast to our previous example, the theoretical literature on network models such as (2.4) is scarce, and to our knowledge no approach has as yet been developed for achieving bias reduction in such a setting.

In this example, a worker's fixed effect may appear in the nuisance parameter η_i

[2]The model relies on the assumption that the network $\mathcal{K}$ of co-workers is exogenous, i.e., independent of the shocks ε_j. Relaxing exogeneity through a parametric model of team formation is conceptually feasible within our likelihood approach, but we do not consider this extension here.

across multiple observations. While such "overlapping fixed effects" typically complicate the analysis and correction of incidental parameter bias, our orthogonalization and estimation methods straightforwardly accommodate this structure.

2.2 The role of first-order orthogonality and its limitations

In the remainder of this section, we motivate our approach in a setting where one wishes to estimate $\mu_0 = \theta_0$ based on a random sample $Z_1, \ldots, Z_n$, taking u to be a univariate function, and η_0 and θ_0 to be scalar. Hence $N = 1$, and $n_1 = n$ is the total number of observations.

If $\mathbb{E}(\sum_{j=1}^n u(Z_j; \theta_0, \eta_0)) = 0$, a conventional estimator of θ_0, say $\widehat{\theta}$, would be the solution to

$$\sum_{j=1}^{n} u(Z_j; \theta, \widehat{\eta}) = 0,$$

where $\widehat{\eta}$ is a consistent estimator of η_0 obtained in a preliminary step. However, it is well known that such a "plug-in" estimator is sensitive to the quality of the preliminary estimator $\widehat{\eta}$ used.

Assuming sufficient regularity, a standard argument based on a linearization around θ_0 yields

$$\left(\mathbb{E}\left(-\frac{\partial u(Z_j; \theta_0, \eta_0)}{\partial \theta}\right) + o_P(1)\right)(\widehat{\theta} - \theta_0) = \frac{1}{n}\sum_{j=1}^{n} u(Z_j; \theta_0, \widehat{\eta}),$$

so that the sampling properties of $\widehat{\theta} - \theta_0$ are dictated by the sampling properties of the estimating equation. We have

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} u(Z_j; \theta_0, \widehat{\eta}) = & \underbrace{\frac{1}{n}\sum_{j=1}^{n} u(Z_j; \theta_0, \eta_0)}_{(A)} \\
& + \underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} \frac{\partial u(Z_j; \theta_0, \eta_0)}{\partial \eta} - \mathbb{E}\left(\frac{\partial u(Z_j; \theta_0, \eta_0)}{\partial \eta}\right)\right)(\widehat{\eta} - \eta_0)}_{(B)} \\
& + \underbrace{\mathbb{E}\left(\frac{\partial u(Z_j; \theta_0, \eta_0)}{\partial \eta}\right)(\widehat{\eta} - \eta_0)}_{(C)} \\
& + O_P(|\widehat{\eta} - \eta_0|^2).
\end{aligned} \tag{2.5}$$

The (A) term in (2.5) is a zero-mean sample average to which a standard central-limit theorem can be applied. Hence, it is generally $O_P(n^{-1/2})$. The next two terms in the expansion capture the first-order effect of estimation noise in $\widehat{\eta}$. The (B) term can generally be ensured to be $o_P(n^{-1/2})$. A generic approach to achieve this is to compute $\widehat{\eta}$ from data that are independent of $Z_1, \ldots, Z_n$, for example using sample splitting. In the case of (2.5), (B) is the product of a sample average of zero-mean random variables—which is $O_P(n^{-1/2})$—and an $o_P(1)$ term— as $\widehat{\eta}$ is consistent for η_0—and, therefore, (B) is $o_P(n^{-1/2})$. The (C) term, however, features a non-random Jacobian that, in general, is non-zero. Hence, (C) is $O_P(|\widehat{\eta} - \eta_0|)$, and will only be asymptotically negligible when $\widehat{\eta}$ is superconsistent for η_0, which is not usually the case.

Suppose now that u is first-order orthogonal, in the sense that

$$\mathbb{E}\left(\frac{\partial u(Z_j; \theta_0, \eta_0)}{\partial \eta}\right) = 0. \tag{2.6}$$

Then the (C) term vanishes from (2.5) and we obtain

$$\frac{1}{n}\sum_{j=1}^{n} u(Z_j; \theta_0, \widehat{\eta}) = \frac{1}{n}\sum_{j=1}^{n} u(Z_j; \theta_0, \eta_0) + O_P(|\widehat{\eta} - \eta_0|^2) + o_P(n^{-1/2}). \tag{2.7}$$

The requirement that $\widehat{\eta} - \eta_0 = o_P(n^{-1/4})$ then guarantees that the impact of the estimation error in $\widehat{\eta}$ on $\widehat{\theta}$ is asymptotically negligible. While a given function u does not, in general, satisfy (2.6), Neyman (1959) proposed a general method to transform it into one that does. The resulting function is said to be Neyman-orthogonal.

Condition (2.6) has a long history in semiparametric estimation problems (Bickel, 1982, Schick, 1986, Newey, 1994). More recently, it has proved to be a fundamental ingredient in the literature on high-dimensional inference (see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018). There are, however, instances where it is ineffective. To illustrate this it suffices to consider the simple panel data setting from the Neyman and Scott (1948) problem.

Example: Neyman-Scott model (continued). In this problem it is easy to verify that

$$\mathbb{E}\left(\frac{\partial u(Y_i; \sigma_0^2, \eta_{i0})}{\partial \eta_i}\right) = -\frac{1}{\sigma_0^4}\sum_{j=1}^{T} \mathbb{E}\left(Y_{ij} - \eta_{i0}\right) = 0,$$

and so the score is already first-order Neyman-orthogonal with respect to the fixed effects. Nevertheless, given preliminary estimators $\widehat{\eta}_1, \ldots, \widehat{\eta}_N$, and letting $\nu_i = \widehat{\eta}_i - \eta_{i0}$, the estimator

$$\widehat{\sigma}^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} (Y_{ij} - \widehat{\eta}_i)^2, \tag{2.8}$$

has expectation $\sigma_0^2 - {}^2/_N \sum_{i=1}^N \mathbb{E}(\bar{\varepsilon}_i \, \nu_i) + {}^1/_N \sum_{i=1}^N \mathbb{E}(\nu_i^2)$, for $\bar{\varepsilon}_i = {}^1/_T \sum_{j=1}^T \varepsilon_{ij}$. Thus, when using sample splitting, the bias is ${}^1/_N \sum_{i=1}^N \mathbb{E}(\nu_i^2)$, the mean squared error of the preliminary estimator. With cross-fitting this is, at best, $O(T^{-1})$. Hence, $\sqrt{NT}(\widehat{\sigma}^2 - \sigma_0^2)$ will not have a correctly-centered limit distribution unless ${}^N/_T \to 0$. However, under this condition, the joint maximum-likelihood estimator of σ_0^2 and the fixed effects, too, is asymptotically unbiased. Hence, having a score that is Neyman-orthogonal, even when combined with sample splitting, does not suffice to resolve the incidental parameter problem in panel data problems.

2.3 Higher-order orthogonality

To see how Neyman-orthogonality to a higher order can be helpful we now consider a further expansion of (2.5). Again assuming sufficient regularity, we have, for any integer $q \geq 1$,

$$\begin{aligned}
\frac{1}{n} \sum_{j=1}^{n} u(Z_j; \theta_0, \widehat{\eta}) = & \underbrace{\frac{1}{n} \sum_{j=1}^{n} u(Z_j; \theta_0, \eta_0)}_{(A)} \\
& + \underbrace{\sum_{p=1}^{q} \frac{1}{p!} \left(\frac{1}{n} \sum_{j=1}^{n} \frac{\partial^p u(Z_j; \theta_0, \eta_0)}{\partial \eta^p} - \mathbb{E}\left(\frac{\partial^p u(Z_j; \theta_0, \eta_0)}{\partial \eta^p} \right) \right) (\widehat{\eta} - \eta_0)^p}_{(B)} \\
& + \underbrace{\sum_{p=1}^{q} \frac{1}{p!} \mathbb{E}\left(\frac{\partial^p u(Z_j; \theta_0, \eta_0)}{\partial \eta^p} \right) (\widehat{\eta} - \eta_0)^p}_{(C)} \\
& + O_P(|\widehat{\eta} - \eta_0|^{q+1}).
\end{aligned}$$

Here, the (A) term is the same as before. Also, with sample splitting we can again ensure that the (B) term will be asymptotically negligible. On the other hand, if the

function u satisfies the higher-order orthogonality condition

$$\mathbb{E}\left(\frac{\partial^p u(Z_j;\theta_0,\eta_0)}{\partial \eta^p}\right) = 0, \qquad 1 \leq p \leq q, \tag{2.9}$$

the (C) term is equal to zero, and so

$$\frac{1}{n}\sum_{j=1}^{n} u(Z_j;\theta_0,\widehat{\eta}) = \frac{1}{n}\sum_{j=1}^{n} u(Z_j;\theta_0,\eta_0) + O_P(|\widehat{\eta}-\eta_0|^{q+1}) + o_P(n^{-1/2}). \tag{2.10}$$

Comparing (2.10) to (2.7) we see that the impact of estimation noise in $\widehat{\eta}$ on our estimator of θ_0 has been reduced further. Moreover, for the impact of estimation error to be negligible, we now only require that $|\widehat{\eta}-\eta_0|^{q+1} = o_P(n^{-1/2})$. It then follows from standard results that, as $n \to \infty$,

$$\sqrt{n}(\widehat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma_\theta)$$

for some Σ_θ, provided that

$$\widehat{\eta}-\eta_0 = o_P(n^{-1/2(q+1)}).$$

The notion of qth-order Neyman-orthogonality as in (2.9) was introduced by Mackey, Syrgkanis and Zadik (2018). In the context of our conditional-likelihood setup, we will give a general procedure to construct higher-order Neyman-orthogonal functions below.

Example: Neyman-Scott model (continued). In the model of Neyman and Scott (1948),

$$\mathbb{E}\left(\frac{\partial^2 u(Y_i;\sigma_0^2,\eta_{i0})}{\partial \eta_i^2}\right) = \frac{T}{\sigma_0^4} \neq 0.$$

It thus follows that u is not orthogonal to second order. Below we will show that a second-order Neyman-orthogonal score equation exists. This estimating equation does not involve the nuisance parameters at all – in particular, it does not depend on the choice of preliminary estimator of η_{i0} – and its solution is the usual degrees-of-

freedom corrected estimator

$$\widehat{\sigma}^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{j=1}^{T} (Y_{ij} - \overline{Y}_i)^2, \tag{2.11}$$

where $\overline{Y}_i = {}^1\!/\!_T \sum_{j=1}^{T} Y_{ij}$. The estimator $\widehat{\sigma}^2$ is well-known to be fixed-T consistent.

3 Estimation based on orthogonalized functions

We now present our estimation approach in the general case where the target parameter μ_0 may be equal to θ_0 or may be a different parameter such as an average effect, and there are multiple, vector-valued nuisance parameters η_{i0}. We start by formally defining higher-order Neyman-orthogonality in this general setup and describe estimation based on higher-order Neyman-orthogonal moment functions. In the next section, we will then show how to construct such functions.

3.1 Definition of higher-order orthogonality

Let d_η be the dimension of η and write $\eta = (\eta_1, \ldots, \eta_{d_\eta})$. For any non-negative integer p and a vector of integers $m = (m_1, \ldots, m_p)$ satisfying $1 \leq m_s \leq d_\eta$ for all $1 \leq s \leq p$, define

$$D_\eta^m = \frac{\partial^p}{\partial \eta_{m_1} \cdots \partial \eta_{m_p}}. \tag{3.1}$$

For a given p, there are $d_p = \binom{d_\eta + p - 1}{p}$ unique such partial derivatives. Let $\nabla_\eta^{(p)}$ be the vector operator of dimension d_p that collects all these unique partial derivatives of order p. Finally, let ∇_η^q be the vector operator of dimension $\sum_{p=1}^{q} d_p$ obtained on stacking $\nabla_\eta^{(p)}$ for $p = 1, \ldots, q$. Explicitly, we have

$$\nabla_\eta^q = \begin{pmatrix} \nabla_\eta^{(1)} \\ \nabla_\eta^{(2)} \\ \vdots \\ \nabla_\eta^{(q)} \end{pmatrix} = \begin{pmatrix} \left[D_\eta^m \,:\, m \in \{1, \ldots, d_\eta\} \right] \\ \left[D_\eta^m \,:\, m \in \{1, \ldots, d_\eta\}^2,\, m_1 \leq m_2 \right] \\ \vdots \\ \left[D_\eta^m \,:\, m \in \{1, \ldots, d_\eta\}^q,\, m_1 \leq m_2 \leq \ldots \leq m_q \right] \end{pmatrix}.$$

Neyman-orthogonality to order q can now be defined as follows (Mackey, Syrgkanis and Zadik, 2018).

Definition 1. *If the function u satisfies*

$$\mathbb{E}\left[\nabla_\eta^q u(Z;\theta_0,\eta_0,\mu_0)\right] = 0, \tag{3.2}$$

for some integer q, then we say that u is Neyman-orthogonal to order q.

In this definition, *all* possible partial derivatives of $u(Z;\theta,\eta,\mu)$ with respect to η up to order q have mean zero.

3.2 Estimation

Let μ_0 satisfy (2.1) for a (possibly vector-valued) function u. We assume that u is Neyman-orthogonal to order q with respect to η_i, in the sense of Definition 1. Suppose that we have access to preliminary estimators $\widehat{\eta}_1,\ldots,\widehat{\eta}_N$ of the nuisance parameters that are independent of the data $Z_1,\ldots,Z_N$. If η_{i0} is defined as the solution to a moment condition involving the same data, estimation based on sample-splitting, combined with cross-fitting (see, e.g., Newey and Robins, 2017), can be applied. When the observations are independent this is conventional. For situations where the data are dependent, modified sample-splitting strategies are available (see, e.g., Semenova, Goldman, Chernozhukov and Taddy, 2023).

We estimate μ_0 by the GMM estimator

$$\widehat{\mu} = \underset{\mu}{\operatorname{argmin}} \left\| \sum_{i=1}^N u(Z_i;\widehat{\theta},\widehat{\eta}_i,\mu) \right\|_W, \tag{3.3}$$

where W is a chosen symmetric positive-definite matrix, $\|u\|_W = \sqrt{u^\top W\, u}$, and $\widehat{\theta}$ is an estimator of θ_0.

The estimator $\widehat{\theta}$ will depend on the problem at hand. If θ_0 is defined through a moment condition of the form $\sum_{i=1}^N \mathbb{E}(\widetilde{u}(Z_i;\theta_0,\eta_{i0})) = 0$, for a function $\widetilde{u}$ that is Neyman-orthogonal to order q, then our framework can be applied and we can use

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\| \sum_{i=1}^N \widetilde{u}(Z_i;\theta,\widehat{\eta}_i) \right\|_{\widetilde{W}}, \tag{3.4}$$

where $\widetilde{W}$ is again a chosen weight matrix. In this case, we may equally combine (3.3) and (3.4) into a single GMM estimation procedure.

In Section 7, we provide conditions under which this approach yields estimators that are $n^{-1/2}$-consistent and asymptotically normal, where n is the total number of observations. We will impose two key conditions. The first one is that, although their number may increase with the sample size, the dimension of each η_i remains bounded as n tends to infinity. This imposes a suitable sense of sparsity in the relationship between the nuisance parameters and the outcomes. This condition is trivially satisfied in the panel data and network problems with fixed effects that we consider. The second key condition we impose is that the convergence rates of the preliminary estimates $\widehat{\eta}_i$ be faster than $n^{-1/2(q+1)}$. This ensures that, after having orthogonalized to order q, any remainder terms are asymptotically negligible.

Our approach requires choosing an orthogonality order q. In practice, we recommend reporting estimates and standard errors for various values $q = 1, 2, ..., q_{\max}$, each based on a q-orthogonal function u_q^*. For a given q that satisfies the conditions of our theory (see Theorem 2 in Section 7), the estimators based on u_q^* and u_{q+1}^* are both root-n consistent and asymptotically normal. Given this, we propose to report as a diagnostic the difference in orthogonal moment functions $u_q^* - u_{q+1}^*$, suitably normalized. A large value of this statistic indicates that the order q of orthogonality may be too small. We describe this approach in Appendix C, while leaving the formal construction of a selection method for q and its impact on inference on μ_0 to future work.

4 Achieving higher-order Neyman-orthogonality

4.1 Main result

Let u be a moment function. We now show how to construct an orthogonalized counterpart of u, which we call u_q^*, that is Neyman-orthogonal to order q, where $q \geq 1$ is any arbitrary order.

Recall the definition of the vector operators $\nabla_\eta^{(p)}$ and ∇_η^q in Subsection 3.1. It is convenient to introduce the Bhattacharyya (1946) basis $v_1, v_2, \ldots$, where

$$v_p(z;\theta,\eta) = \frac{\nabla_\eta^{(p)} \ell(y \,|\, x;\theta,\eta)}{\ell(y \,|\, x;\theta,\eta)}.$$

McLeish and Small (1994) discuss several properties of this basis. One important

property for our purposes is that

$$\mathbb{E}_{\theta,\eta}(v_p(Z;\theta,\eta)\,|\,X=x)=\int v_p(z;\theta,\eta)\,\ell(y\,|\,x;\theta,\eta)\,dy=0 \tag{4.1}$$

for any p, so all elements of the Bhattacharyya basis have (conditional) mean equal to zero. In (4.1), and throughout this section, $\mathbb{E}_{\theta,\eta}(\cdot\,|\,X=x)$ denotes the conditional expectation under $\ell(y\,|\,x;\theta,\eta)$.

The low-order basis functions are familiar from likelihood theory. For example,

$$\begin{aligned} v_1(z;\theta,\eta) &= \frac{\partial \log \ell(y\,|\,x;\theta,\eta)}{\partial \eta}, \\ v_2(z;\theta,\eta) &= \frac{\partial \log \ell(y\,|\,x;\theta,\eta)}{\partial \eta}\frac{\partial \log \ell(y\,|\,x;\theta,\eta)}{\partial \eta^\top}+\frac{\partial^2 \log \ell(y\,|\,x;\theta,\eta)}{\partial \eta \partial \eta^\top}. \end{aligned}$$

The fact that these functions have mean zero follows from the unbiasedness of the score and from the information equality, respectively.

Stacking the leading q basis functions, we obtain

$$w_q(z;\theta,\eta)=\frac{\nabla_\eta^q \ell(y\,|\,x;\theta,\eta)}{\ell(y\,|\,x;\theta,\eta)}=\begin{pmatrix} v_1(z;\theta,\eta) \\ \vdots \\ v_q(z;\theta,\eta) \end{pmatrix}.$$

The vectors w_q are mean-zero "generalized score functions". The vector space spanned by the Bhattacharyya basis at order q is a higher-order tangent set.[3] While it is possible to achieve higher-order Neyman-orthogonality using other bases of functions, the Bhattacharyya basis delivers simple expressions through the use of Bartlett identities.

Next, let us define the matrices

$$\Sigma_{w_q w_q}(x;\theta,\eta)=\mathbb{E}_{\theta,\eta}(w_q(Z;\theta,\eta)\,w_q(Z;\theta,\eta)^\top\,|\,X=x),$$

and

$$\Sigma_{w_q u}(x;\theta,\eta,\mu)=\mathbb{E}_{\theta,\eta}(w_q(Z;\theta,\eta)\,u(Z;\theta,\eta,\mu)^\top\,|\,X=x),$$

which are, respectively, the (conditional) covariance matrix of the first q members

[3] In models with i.i.d. observations, for which the likelihood of any q observations is $\prod_{r=1}^q \ell(z_r)$, the tangent set of order q is defined as the span of all derivatives with respect to the parameter, up to order q, of $\prod_{r=1}^q \ell(z_r)$, divided by $\prod_{r=1}^q \ell(z_r)$; see equation (1.8) in van der Vaart (2014). In our setting we do not assume that observations are i.i.d.

of the Bhattacharyya basis, and the covariance matrix of the same q basis functions with the vector function u. Finally, let

$$b_q(x;\theta,\eta,\mu) = \nabla_\eta^q \, \mathbb{E}_{\theta,\eta}(u(Z;\theta,\eta,\mu)^\top | X = x).$$

Note that b_q is zero when u is the score for θ, i.e., $\frac{\partial \log \ell(y \,|\, x;\theta,\eta)}{\partial \theta}$. In general, however, b_q will be non-zero. Here we assume that $u(z;\theta,\eta,\mu)$ and $\ell(y \,|\, x;\theta,\eta)$ are sufficiently often differentiable in η, and that the expectations in the definitions of $\Sigma_{w_q w_q}$, $\Sigma_{w_q u}$, and b_q are well-defined.

The proof of the following result is in Appendix A.

Theorem 1. *Suppose that $\Sigma_{w_q w_q}(x;\theta,\eta)$ is invertible and let*

$$A(x;\theta,\eta,\mu) = \Sigma_{w_q w_q}(x;\theta,\eta)^{-1} \left(\Sigma_{w_q u}(x;\theta,\eta,\mu) - b_q(x;\theta,\eta,\mu) \right).$$

Then the function

$$u_q^*(z;\theta,\eta,\mu) = u(z;\theta,\eta,\mu) - A(x;\theta,\eta,\mu)^\top \, w_q(z;\theta,\eta)$$

satisfies $\mathbb{E}_{\theta,\eta}(\nabla_\eta^q \, u_q^*(Z;\theta,\eta,\mu) \,|\, X = x) = 0$. This implies that u_q^* is Neyman-orthogonal to order q, as defined above.

Theorem 1 generalizes the projected-score construction of Small and McLeish (1989) and Waterman and Lindsay (1996). To see this, consider the case where $\mu_0 = \theta_0$, and u is the score function for θ. Then $b_q = 0$ and Theorem 1 yields

$$u_q^*(z;\theta,\eta,\mu) = u(z;\theta,\eta,\mu) - \left(\Sigma_{w_q w_q}(x;\theta,\eta)^{-1} \Sigma_{w_q u}(x;\theta,\eta) \right)^\top \, w_q(z;\theta,\eta),$$

which is the projected score of order q. The projected score was originally developed as a tool to achieve E-ancillarity (Small and McLeish, 1988) and to approximate the conditional score for θ, when the latter exists (Waterman and Lindsay, 1996). It generalizes Neyman (1959) in that u_q^* is the (population) residual of a least-squares regression of u on w_q; thus,

$$\mathbb{E}_{\theta,\eta}(w_q(Z;\theta,\eta) \, u_q^*(Z;\theta,\eta,\mu)^\top | X = x) = 0.$$

While the fact that u_q^* is Neyman-orthogonal is noted by Waterman and Lindsay

(1996) (although a link with Neyman's work is not made), it is not exploited. Moreover, unlike the conditional score, the projected score still depends on η, and it will generally not have improved properties over the score itself. As we highlight here, it is the combination of higher-order versions of Neyman-orthogonality with sample splitting that allows one to improve over working with the original score.

Theorem 1 covers more general estimating equations as well as more general parameters of interest, such as average elasticities or counterfactual quantities. Incorporating $b_q(x;\theta,\eta,\mu)$ into $u_q^*(z;\theta,\eta,\mu)$ is a key innovation that enables this generality. Observe that, in this case, we have that

$$\mathbb{E}_{\theta,\eta}(w_q(Z;\theta,\eta)\, u_q^*(Z;\theta,\eta,\mu)^\top | X = x) = b_q(x;\theta,\eta,\mu),$$

thereby revealing u_q^* to be an influence function of order q per Equation (1.9) in van der Vaart (2014).

We note that Theorem 1 requires the matrix $\Sigma_{w_q w_q}(x;\theta,\eta)$ to be invertible. In the standard case of first-order Neyman-orthogonality this corresponds to non-singularity of the information matrix of the nuisance parameters. For higher-order Neyman-orthogonality this requirement imposes further restrictions. For example, in the team production model (2.4) with two-worker teams, the 2×2 matrix $\Sigma_{w_1 w_1}(\mathcal{K};\theta,\eta)$ is not invertible, as only the sum $\eta_{k(j,1)}^\gamma + \eta_{k(j,2)}^\gamma$ can be identified. In our application in Section 6, we will tackle this issue by combining data on teams of size 2 with single-author articles, and working with subsets i of three teams each.[4]

4.2 Intuition and discussion

To gain intuition into the construction in Theorem 1 it is useful to again consider the case where u is a univariate function, the nuisance parameter is a scalar, and one wishes to estimate $\mu_0 = \theta_0$. Throughout this subsection we also assume that the conditional mean $\mathbb{E}_{\theta,\eta}\left(u(Z;\theta,\eta) \,|\, X = x\right)$ does not vary with η, so that the term $b_q(x;\theta,\eta,\mu)$ in Theorem 1 is zero; this holds, for example, when u is the score for θ. The general case is analogous, with the b_q term retained.

[4] As another example, consider the standard binary-choice panel data model

$$\mathbb{P}_{\theta,\eta_i}(Y_{ij} = 1 \,|\, X_{i1}, ..., X_{iT}) = \Phi(\eta_i + X_{ij}^\top \theta), \quad i = 1, \ldots, N, \quad j = 1, \ldots, T,$$

for (conditionally-independent) binary outcomes Y_{ij} and covariates X_{ij}. In this model, the rank of $\Sigma_{w_q w_q}(x;\theta,\eta)$ is bounded by $2^T - 1$, so $\Sigma_{w_q w_q}(x;\theta,\eta)$ is singular for all $q \geq 2^T$.

First-order orthogonality. To relate our approach to the literature consider first $q = 1$. Let

$$u_1^*(z;\theta,\eta) = u(z;\theta,\eta) - a_1(x;\theta,\eta)\, v_1(z;\theta,\eta), \tag{4.2}$$

for some function a_1. Note that, by virtue of (4.1), the term involving v_1 does not introduce any bias. We have

$$\frac{\partial u_1^*(z;\theta,\eta)}{\partial \eta} = \frac{\partial u(z;\theta,\eta)}{\partial \eta} - \frac{\partial a_1(x;\theta,\eta)}{\partial \eta}\, v_1(z;\theta,\eta) - a_1(x;\theta,\eta)\, \frac{\partial v_1(z;\theta,\eta)}{\partial \eta}.$$

Take conditional expectations and exploit (4.1) to see that

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial u_1^*(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = 0$$

if and only if

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial u(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) - a_1(x;\theta,\eta)\ \mathbb{E}_{\theta,\eta}\left(\frac{\partial v_1(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = 0.$$

This is achieved by setting

$$a_1(x;\theta,\eta) = \left(\mathbb{E}_{\theta,\eta}\left(\frac{\partial v_1(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) \right)^{-1} \mathbb{E}_{\theta,\eta}\left(\frac{\partial u(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right). \tag{4.3}$$

Iterating expectations shows that the resulting function u_1^* is Neyman-orthogonal to order $q = 1$. By the information matrix equality we have

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial v_1(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = -\mathbb{E}_{\theta,\eta}\left(v_1(Z;\theta,\eta)^2 \middle| X = x \right) = -\Sigma_{w_1 w_1}(x;\theta,\eta),$$

and

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial u(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = -\mathbb{E}_{\theta,\eta}\left(v_1(Z;\theta,\eta) u(Z;\theta,\eta) \middle| X = x \right) = -\Sigma_{w_1 u}(x;\theta,\eta),$$

leading to the representation of the function u_1^* as in the theorem.

The above derivation of (4.3) is well-known. Furthermore, it does not hinge on the likelihood structure. Indeed, recent work exploiting orthogonality, such as that surveyed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins

(2018), does so in the context of moment conditions. In our setup, as in Neyman's (1959) original work, the likelihood structure implies that a_1 is known up to the model parameters θ and η (conditional on the regressors). Outside of this framework, in contrast, a_1 needs to be treated as an additional nuisance parameter. This is possible because, as u_1^* is linear in a_1, it is automatically first-order Neyman-orthogonal to it by virtue of (4.1). This logic, however, does not extend to higher order, as the implied system of equations becomes inconsistent, so that no solution exists, as we will see next.

Higher-order orthogonality. Let $q = 2$, and again consider a linear transformation of u, now involving the leading two Bhattacharyya basis functions. This gives

$$u_2^*(z;\theta,\eta) = u(z;\theta,\eta) - \begin{pmatrix} a_{21}(x;\theta,\eta) \\ a_{22}(x;\theta,\eta) \end{pmatrix}^\top \begin{pmatrix} v_1(z;\theta,\eta) \\ v_2(z;\theta,\eta) \end{pmatrix}. \tag{4.4}$$

Taking first-derivatives with respect to the nuisance parameter, and proceeding as in the first-order case, gives

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial u(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = \begin{pmatrix} a_{21}(x;\theta,\eta) \\ a_{22}(x;\theta,\eta) \end{pmatrix}^\top \begin{pmatrix} \mathbb{E}_{\theta,\eta}\left(\frac{\partial v_1(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) \\ \mathbb{E}_{\theta,\eta}\left(\frac{\partial v_2(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) \end{pmatrix}.$$

Solving this equation for a_{21} for given a_{22} yields

$$a_{21}(x;\theta,\eta) = a_1(x;\theta,\eta) - c_1(x;\theta,\eta)\, a_{22}(x;\theta,\eta), \tag{4.5}$$

where a_1 is given by (4.3) and

$$c_1(x;\theta,\eta) = \left(\mathbb{E}_{\theta,\eta}\left(\frac{\partial v_1(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) \right)^{-1} \mathbb{E}_{\theta,\eta}\left(\frac{\partial v_2(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right).$$

The coefficient c_1 has the same form as a_1, except that it features v_2 instead of u. Moreover, plugging (4.5) back into (4.4) yields

$$u_2^*(z;\theta,\eta) = u_1^*(z;\theta,\eta) - a_{22}(x;\theta,\eta)\, v_2^*(z;\theta,\eta),$$

where $v_2^*(z;\theta,\eta) = v_2(z;\theta,\eta) - c_1(x;\theta,\eta)\, v_1(z;\theta,\eta)$. Note that v_2^* is Neyman-orthogonal to order 1, that is,

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial v_2^*(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = 0.$$

It follows that u_2^* is Neyman-orthogonal to order 1 for any a_{22}. We will now choose a_{22} such that u_2^* is Neyman-orthogonal to order 2.

Next, differentiating u_2^* with respect to η twice gives

$$\begin{aligned}\frac{\partial^2 u_2^*(z;\theta,\eta)}{\partial \eta^2} = \frac{\partial^2 u_1^*(z;\theta,\eta)}{\partial \eta^2} - a_{22}(x;\theta,\eta)\, \frac{\partial^2 v_2^*(z;\theta,\eta)}{\partial \eta^2} \\ - \frac{\partial^2 a_{22}(x;\theta,\eta)}{\partial \eta^2}\, v_2^*(z;\theta,\eta) - 2\frac{\partial a_{22}(x;\theta,\eta)}{\partial \eta}\, \frac{\partial v_2^*(z;\theta,\eta)}{\partial \eta}.\end{aligned}$$

Since v_2^* has zero mean and is orthogonal to order 1, the terms involving the first and second derivative of a_{22} drop out when taking expectations. It follows that u_2^* in (4.4) is Neyman-orthogonal to order 2 when one sets a_{21} to its expression in (4.5), and a_{22} to

$$a_{22}(x;\theta,\eta) = \left(\mathbb{E}_{\theta,\eta}\left(\frac{\partial^2 v_2^*(Z;\theta,\eta)}{\partial \eta^2} \middle| X = x \right) \right)^{-1} \mathbb{E}_{\theta,\eta}\left(\frac{\partial^2 u_1^*(Z;\theta,\eta)}{\partial \eta^2} \middle| X = x \right). \quad (4.6)$$

Note that this construction amounts to solving a system of linear equations. The fact that the solution in (4.5)–(4.6) coincides with the expression in Theorem 1 may then again be verified by using Bartlett identities.

To appreciate the role of the likelihood structure in the above argument, suppose that a_{21} and a_{22} are not known up to the parameters θ and η. Then they are additional nuisance parameters and, thus, we require that all first- and second-order derivatives with respect to (a_{21}, a_{22}), and η be mean zero. The cross-derivatives between (a_{21}, a_{22}) and η are problematic, since having those to be mean zero would require

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial v_1(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = 0, \qquad \mathbb{E}_{\theta,\eta}\left(\frac{\partial v_2(Z;\theta,\eta)}{\partial \eta} \middle| X = x \right) = 0,$$

which is not generally the case.

5 Examples

5.1 Panel data models

Consider an $N \times T$ panel data model with individual effects. Here, the likelihood factors across the cross-sectional observations and the likelihood contribution of unit i takes the form

$$\prod_{j=1}^{T} f(Y_{ij} \,|\, X_{ij}; \theta, \eta_i).$$

The maximum-likelihood estimator is well-known to suffer from a bias that is $O(T^{-1})$; see Hahn and Newey (2004) and Hahn and Kuersteiner (2011) for derivations of this bias in static and dynamic models, respectively. Consider the estimation of θ_0. The bias in the estimator comes from bias in the score stemming from estimation noise in the fixed effects. Taking η_i to be scalar for notational simplicity, and letting $\widehat{\eta}_i$ be an estimator of η_{i0}, an expansion of the (normalized) score[5]

$$u(Z_i; \theta_0, \widehat{\eta}_i) = \frac{1}{T} \sum_{j=1}^{T} \frac{\partial \log f(Y_{ij}|X_{ij}; \theta_0, \widehat{\eta}_i)}{\partial \theta}$$

yields

$$\begin{aligned} u(Z_i; \theta_0, \widehat{\eta}_i) =& u(Z_i; \theta_0, \eta_{i0}) + \frac{\partial u(Z_i; \theta_0, \eta_{i0})}{\partial \eta_i} (\widehat{\eta}_i - \eta_{i0}) + \frac{1}{2} \frac{\partial^2 u(Z_i; \theta_0, \eta_{i0})}{\partial \eta_i^2} (\widehat{\eta}_i - \eta_{i0})^2 \\ &+ o_P(|\widehat{\eta}_i - \eta_{i0}|^2). \end{aligned}$$

Taking expectations and re-arranging shows that

$$\begin{aligned} \mathbb{E}(u(Z_i; \theta_0, \widehat{\eta}_i)) =& \operatorname{cov}\left(\frac{\partial u(Z_i; \theta_0, \eta_{i0})}{\partial \eta_i}, \widehat{\eta}_i - \eta_{i0} \right) \\ &+ \mathbb{E}\left(\frac{\partial u(Z_i; \theta_0, \eta_{i0})}{\partial \eta_i} \right) \mathbb{E}(\widehat{\eta}_i - \eta_{i0}) \\ &+ \frac{1}{2} \mathbb{E}\left(\frac{\partial^2 u(Z_i; \theta_0, \eta_{i0})}{\partial \eta_i^2} \right) \mathbb{E}((\widehat{\eta}_i - \eta_{i0})^2) + o(\mathbb{E}(|\widehat{\eta}_i - \eta_{i0}|^2)). \end{aligned}$$

[5] In this discussion we work with the score divided by T, to facilitate the comparison with the panel data literature.

If we set $\widehat{\eta}_i = \widehat{\eta}_i(\theta_0) = \arg\max_\eta \sum_{j=1}^T \log f(Y_{ij} \,|\, X_{ij}; \theta_0, \eta)$, the maximum-likelihood estimator (MLE) given θ_0, then each one of these terms is $O(T^{-1})$. If we use an estimator $\widehat{\eta}_i$ that is independent of the estimation sample, the first term disappears. However, the remaining terms, which capture the nonlinearity bias and variance in the estimator of η_{i0}, remain. Hahn and Newey (2004), Arellano and Hahn (2007), and Dhaene and Jochmans (2015a,b) present estimators of these terms based on the MLE that can be used to construct a bias-corrected estimator.

Lancaster (2002) and Woutersen (2002) integrate-out the fixed effects using a uniform prior after orthogonalizing to order 1 to obtain an estimator with bias $o(T^{-1})$; Arellano (2003) presents an alternative derivation of the same result. First-order Neyman-orthogonality, by itself, does not suffice as it does not handle the third term in the expansion, that is, it does not properly correct for the noise in the estimated fixed effects. Li, Lindsay and Waterman (2003), building on Waterman and Lindsay (1996), show that their (second-order) projected score for θ, when evaluated at $\widehat{\eta}_i(\theta)$, is a first-order unbiased estimating equation for θ. Thus, here, a sample-splitting procedure is not needed to achieve bias reduction. This is a consequence of the (second- or higher-order) projected score being orthogonal to the influence function of $\widehat{\eta}_i(\theta)$. While interesting, it is not clear whether this property extends to higher-order projections or to other parameters of interest, such as average marginal effects.

More generally, with $\widehat{\eta}_i - \eta_{i0} = O_P(T^{-1/2})$, the score admits a higher-order expansion of the form,

$$\mathbb{E}(u(Z_i; \theta_0, \widehat{\eta}_i)) = \frac{B_1}{T} + \frac{B_2}{T^2} + \cdots + \frac{B_q}{T^q} + o(T^{-q})$$

for constants $B_1, B_2, \ldots, B_q$. The maximum-likelihood estimator has $B_1 \neq 0$, in general, and so requires that $N/T \to 0$ to be asymptotically unbiased. The approaches to bias correction mentioned above remove B_1 but not the remaining terms. Approaches that estimate and subsequently remove all B_p, $1 \leq p \leq q$, to enable asymptotically valid inference are given by Dhaene and Jochmans (2015a,b). Relatedly, an estimator based on Neyman-orthogonalization, combined with a sample-splitting estimator that uses preliminary estimators that satisfy $\widehat{\eta}_i - \eta_{i0} = O_P(T^{-1/2})$, renders the asymptotic bias negligible provided $N = o(T^q)$. The two approaches are not equivalent, however. Removing $B_1, \ldots, B_q$ leaves a bias of order $o(T^{-q})$, which permits asymptotically-valid inference under the weaker requirement that $N = o(T^{2q-1})$. Our rate condition

$N = o(T^q)$, which follows from Assumption 1(v) below, is more restrictive as soon as $q \geq 2$. In exchange, our approach does not require characterizing, estimating, or removing the individual bias terms B_p.

Example: Neyman-Scott model (continued). Recall that the (un-normalized) unit-specific score for σ^2 is given by (2.3). The leading two elements of the Bhattacharyya basis for η_i are

$$v_1(Y_i; \sigma^2, \eta_i) = \sum_{j=1}^{T} \frac{Y_{ij} - \eta_i}{\sigma^2}, \qquad v_2(Y_i; \sigma^2, \eta_i) = -\frac{T}{\sigma^2} + \left(\sum_{j=1}^{T} \frac{Y_{ij} - \eta_i}{\sigma^2} \right)^2.$$

We apply Theorem 1. A small calculation yields $A(\sigma^2, \eta_i) = (0, 1/2T)^\top$ and, after re-arranging,

$$u_2^*(Y_i; \sigma^2, \eta_i) = \frac{1}{2\sigma^2} \left(\frac{\sum_{j=1}^{T} (Y_{ij} - \overline{Y}_i)^2}{\sigma^2} - (T-1) \right),$$

which does not depend on η_i. Summing over the cross-sectional units gives the second-order orthogonalized score equation for σ^2 as

$$\sum_{i=1}^{N} u_2^*(Y_i; \sigma^2, \eta_i) = \frac{1}{2\sigma^2} \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{T} (Y_{ij} - \overline{Y}_i)^2}{\sigma^2} - N(T-1) \right) = 0,$$

which yields the degrees-of-freedom corrected estimator $\widehat{\sigma}^2$ in (2.11).

Another parameter of interest in this problem is $\mu = 1/N \sum_{i=1}^{N} h(\eta_i)$, for h a known function. This fits our framework, with

$$u(Y_i; \sigma^2, \eta_i, \mu) = h(\eta_i) - \mu.$$

The q-th orthogonalized counterpart to u given by Theorem 1 is available in closed form, as

$$u_q^*(Y_i; \sigma^2, \eta_i, \mu) = h(\eta_i) - \mu + \sum_{k=1}^{q} \sigma^k T^{-\frac{k}{2}} \frac{1}{k!} \nabla_\eta^{(k)} h(\eta_i) H_k \left(\frac{\sum_{j=1}^{T} (Y_{ij} - \eta_i)}{\sqrt{T}\sigma} \right), \tag{5.1}$$

where H_k denotes the k-th Hermite polynomial.

Interestingly, (5.1) shows that, depending on the form of h, the variance of u_q^*

may converge or diverge as q tends to infinity. Indeed, using a property of Hermite polynomials, the variance is

$$\operatorname{Var}\left(u_q^*(Y_i;\sigma^2,\eta_i,\mu)\right) = \sum_{k=1}^{q} \sigma^{2k} T^{-k} \frac{1}{k!} \left[\nabla_\eta^{(k)} h(\eta_i)\right]^2.$$

It is instructive to consider the following three cases: (1) when $h(\eta_i)$ is a polynomial of order K the series is stationary for $q \geq K$; (2) when $h(\eta_i) = \exp(\eta_i)$ the series converges; (3) in contrast, when $h(\eta_i) = \log(\eta_i)$ the series diverges. In the first two cases, using a large q reduces bias without causing variance to diverge, while the third case presents a sharp trade-off between reduced bias and exploding variance as q increases.

5.2 Nonlinear network regression

Our next example is the nonlinear regression model with $d \geq 1$ outcomes,

$$Y_i = m(X_i;\theta_0,\eta_{i0}) + \sigma(X_i;\theta_0)\varepsilon_i, \quad \varepsilon_i \,|\, X \sim \text{iid } \mathcal{N}(0, I_d), \tag{5.2}$$

where $m(x;\theta,\eta_i)$ is a $d \times 1$ vector, $\sigma(x;\theta)$ is a $d \times d$ diagonal matrix, and m and σ are known functions. We will show below that our CES production function example, in logarithms, fits into this framework.

For this model there are no analytical solutions for the orthogonalized estimators. We thus proceed numerically. To construct Neyman-orthogonal moment functions according to Theorem 1 we need to compute $\Sigma_{w_q w_q}(x;\theta,\eta)$, $\Sigma_{w_q u}(x;\theta,\eta,\mu)$, and $b_q(x;\theta,\eta,\mu)$, which involve higher-order derivatives of the conditional likelihood. To compute these derivatives, it is convenient to introduce the operator ∇_m^q that collects all derivatives with respect to the d-vector m up to order q. By the chain rule,

$$\nabla_{\eta_i}^q \ell(y \,|\, x;\theta,\eta_i) = M(x,\theta,\eta_i) \nabla_m^q \ell(y \,|\, x;\theta,\eta_i),$$

where the matrix M has an analytical expression given by the multivariate Faà di Bruno formula (Constantine and Savits, 1996). Given the matrix M it is easy to

compute $\Sigma_{w_q w_q}$, $\Sigma_{w_q u}$, and b_q. For example,

$$\begin{aligned}&\Sigma_{w_q w_q}(x;\theta,\eta_i)\\&= M(x,\theta,\eta_i)\,\mathbb{E}_{\theta,\eta_i}\left(\frac{\nabla_m^q \ell(Y_i \,|\, X_i;\theta,\eta_i)}{\ell(Y_i \,|\, X_i;\theta,\eta_i)}\,\frac{\nabla_m^q \ell(Y_i \,|\, X_i;\theta,\eta_i)}{\ell(Y_i \,|\, X_i;\theta,\eta_i)}^\top \,\middle|\, X_i = x\right) M(x,\theta,\eta_i)^\top,\end{aligned}$$

where the expectation on the right-hand side can be readily computed by relying on formulas for moments of Hermite polynomials. We relegate further details to Appendix D. In the next section we present simulations and an empirical application based on a version of (5.2) designed to study team production.

Example: CES production function (continued). Consider the team production model

$$Y_j = \beta_0(s_j)\left(\frac{1}{s_j}\sum_{r=1}^{s_j}\eta_{k(j,r)0}^{\gamma_0(s_j)}\right)^{\frac{1}{\gamma_0(s_j)}}\varepsilon_j^{\sigma_0(s_j)}, \quad \log\varepsilon_j \,|\, \mathcal{K} \sim \text{iid}\,\mathcal{N}(0,1)\,, \tag{5.3}$$

where s_j is the size of team $j = 1,...,n$, $(k(j,1),\ldots,k(j,s_j))$ are the s_j workers in team j, and the set $\mathcal{K} = \{(k(j,1),\ldots,k(j,s_j)) : j = 1,\ldots,n\}$ collects the workers in all teams. Model (5.3) generalizes Model (2.4) by allowing for teams of varying sizes. Here we focus on teams of size 1 and 2, as in our application, and impose the normalization $\beta_0(1) = 1$. For simplicity we will denote $\beta_0 = \beta_0(2)$ and $\gamma_0 = \gamma_0(2)$, which are the team size and substitution parameters, respectively, in teams of size 2.

We now explain how (5.3) can be written as a special case of (5.2), for a suitable choice of subsets of observations. To any team j of size 2 involving workers k and k', we associate a team $j_1(j)$ of size 1 only involving worker k, and a team $j_2(j)$ of size 1 only involving worker k'. This construction results in N subsets of three teams each. We then write the outcomes for these three teams, in logarithms, as

$$\log Y_j = \log\beta_0 + \frac{1}{\gamma_0}\log\left(\frac{\eta_{k(j,1)0}^{\gamma_0} + \eta_{k(j,2)0}^{\gamma_0}}{2}\right) + \sigma_0(2)\log\varepsilon_j, \tag{5.4}$$

$$\log Y_{j_1(j)} = \log\eta_{k(j,1)0} + \sigma_0(1)\log\varepsilon_{j_1(j)}, \tag{5.5}$$

$$\log Y_{j_2(j)} = \log\eta_{k(j,2)0} + \sigma_0(1)\log\varepsilon_{j_2(j)}, \tag{5.6}$$

which takes the same form as (5.2), for $d = 3$, $\theta_0 = (\beta_0, \gamma_0, \sigma_0^2(1), \sigma_0^2(2))^\top$, Y_i the

vector of the three outcomes in (5.4)–(5.6) for subset i, and η_{i0} the 2×1 vector of worker-specific effects in the corresponding teams.

Remark 1. (Implementation in other models) In models with discrete outcomes Y, one can express the matrices that feature in the expression for $A(x;\theta,\eta,\mu)$ in Theorem 1 in closed form, as sums over the support of Y. In models with continuous outcomes, one can proceed by simulation as follows, in the spirit of the "reparameterization trick" (Kingma and Welling, 2014). Write $Y = g(X,U;\theta,\eta)$ where $U \,|\, X \sim F_U$ (for example, a standard multivariate Gaussian). Let $U^{(s)}$, $s = 1,...,S$, be i.i.d. draws from F_U, and let $Y^{(s)} = g(x,U^{(s)};\theta,\eta)$ and $Z^{(s)} = (Y^{(s)},x)$. Assuming that g is a smooth function of η one can construct the simulation-based counterpart

$$\widehat{A}(x;\theta,\eta,\mu) = \left(\sum_{s=1}^{S} w_q(Z^{(s)};\theta,\eta)w_q(Z^{(s)};\theta,\eta)^\top\right)^{-1} \times \left[\sum_{s=1}^{S} w_q(Z^{(s)};\theta,\eta)u(Z^{(s)};\theta,\eta,\mu)^\top - \sum_{s=1}^{S} \nabla_\eta^q u\left(g(x,U^{(s)};\theta,\eta),x;\theta,\eta,\mu\right)^\top\right].$$

When focusing on θ_0 instead of μ_0, one can rely on the simpler expression

$$\widehat{A}(x;\theta,\eta,\mu) = \left(\sum_{s=1}^{S} w_q(Z^{(s)};\theta,\eta)w_q(Z^{(s)};\theta,\eta)^\top\right)^{-1} \sum_{s=1}^{S} w_q(Z^{(s)};\theta,\eta)u(Z^{(s)};\theta,\eta)^\top,$$

and u_q^* can be obtained by regressing $u(Z^{(s)};\theta,\eta)$ on $w_q(Z^{(s)};\theta,\eta)$. We leave the study of the impact of a finite number S of draws on inference about μ_0 and θ_0 to future work.

6 Application to team production

6.1 Model, data, and implementation

We wish to estimate the parameters of the team production model in (5.4)–(5.6). We will be especially interested in estimating the substitution parameter γ, which drives the nature of complementarities in teams of size 2, and the team size parameter β, which reflects the premium (or penalty) associated with working together relative to working alone. In addition to estimating production-function parameters, we will also report estimates of a counterfactual random re-allocation of workers to teams. Under

random assignment, average output in teams of size 2 can be written as

$$\mathbb{E}^{\text{rand}}(Y_j) = \frac{2}{n_2(n_2-1)} \sum_{k_1<k_2} \beta_0 \left[\frac{1}{2}\left({\eta_{k_1 0}}^{\gamma_0} + {\eta_{k_2 0}}^{\gamma_0}\right)\right]^{\frac{1}{\gamma_0}} \exp\left(\frac{1}{2}\sigma_0^2(2)\right), \tag{6.1}$$

where n_2 denotes the number of authors in teams of size 2. As this quantity is an average over the worker fixed effects, it can be orthogonalized with respect to them using our approach.

Ahmadpoor and Jones (2019) consider model (5.3) without the error term ε_j. Here our goal is to address the statistical challenge caused by the presence of a large number of possibly imprecisely estimated fixed effects. An alternative would be to specify a distribution for author heterogeneity conditional on the team network (i.e., for all the η_{i0}'s conditional on $\mathcal{K}$), as in Bonhomme (2021). An advantage of such a procedure is that, under correct specification, estimates are consistent even in poorly connected networks. This random-effect approach requires, however, to model how authors sort and collaborate in teams. Our approach avoids the need to do so. On the other hand, a fixed-effect approach requires that the author effects can be consistently estimated. In less well-connected networks, the convergence rate will be slower. Orthogonalization to a higher-order allows us to reduce the impact of estimation noise.

We look at the production of academic work in economics. We use data from Ductor, Fafchamps, Goyal and Van der Leij (2014), drawn from the EconLit database. These data contain a large collection of articles, indicated by their ID, together with author identifiers and a measure of journal quality proposed by Kodrzycki and Yu (2006). This measure is a ranking between 0 and 100, which we net of multiplicative time effects and will use as our outcome variable. In order to mitigate the variation in author fixed effects over time while ensuring sufficiently many collaborations, we restrict the sample to articles published between 1990 and 1999, written either alone or with a single co-author.[6] We only include authors who produced at least two sole-authored articles during the sampling period.

Our sample contains 91,626 articles, 10% of which are co-authored, and 16,408 authors. Average journal quality differs greatly across authors, with the 10th percentile

[6] We have also estimated the model on the entire sample, which ranges from 1970 to 1999. While the estimates of the substitution parameter differ somewhat in this larger sample, they are also less stable due to the fact that connectivity is lower. The estimates of the other parameters are similar to the ones on the 1990-1999 subsample.

of the quality measure being 0.4, the median being 0.9, and the 90th percentile being 8.5. The between-author variance in journal quality is 42% of the overall variance. The distribution of journal quality, in turn, is skewed to the right, with a median of 0.6, a 90th percentile of 12, and a 99th percentile of 52. The number of publications per author varies substantially, with a 10th percentile of 2, a median of 4, and a 90th percentile of 13.

To implement our approach, we construct subsets of three papers, one co-authored (j) and two sole-authored $(j_1(j), j_2(j))$, as described in (5.4)–(5.6). The score for θ based on subset i then involves the three teams j, $j_1(j)$, and $j_2(j)$. Proceeding in this way is helpful as it limits the dimension of the parameter η_i to two. This is not only in line with the assumptions we make in deriving asymptotics, but also helpful in terms of computation. Moreover, it reduces the number of derivatives that need to be computed. The number of derivatives nevertheless remains substantial, as we need to compute 9 derivatives at order 2, 19 at order 3, and 55 at order 5, for example; these are the derivatives with respect to the three-dimensional mean vector m in (5.2), from which the derivatives with respect to η_i are obtained by the chain rule. Yet, using the computational remarks from Subsection 5.2, this remains practical.

Finally, we exploit the network structure of the data to perform our sample splitting. For every worker, we construct a preliminary estimator of her fixed effect (in logs) as the average quality of her single-authored papers, except for one that we select at random and use later in estimation. This strategy is feasible due to our sample restriction. For each subset i of three teams, we then stack the two worker fixed effects together to form our preliminary estimate $\widehat{\eta}_i$. We next estimate the parameters $\beta_0, \gamma_0, \sigma_0^2(1), \sigma_0^2(2)$ on the sample from which all these single-authored articles have been removed. In the present case, $\widetilde{u}$ in (3.4) has four components that correspond to the score with respect to all the parameters, and the weight matrix $\widetilde{W}$ is irrelevant since the problem is just-identified. In order to limit the variability due to the choice of split, we average parameter estimates across 100 random splits, through cross-fitting. The bias in the parameter estimates takes a complex form due to the team network environment. In Appendix E we assess the ability of our orthogonalization approach to alleviate this bias in a Monte Carlo simulation.

Table 1: Estimation results

	Substitution γ	Team size β	Variance $\sigma^2(2)$	Variance $\sigma^2(1)$	Diagnostic
Plug-in	0.1267	1.2893	1.6232	1.6395	-
	(0.0386)	(0.0222)	(0.0218)	(0.0274)	
$q = 1$	-1.7279	1.3238	1.6775	1.7330	592.66
	(0.1758)	(0.0267)	(0.0255)	(0.0247)	
$q = 2$	0.7673	1.2956	1.4339	1.4683	195.77
	(0.1939)	(0.0305)	(0.0241)	(0.0262)	
$q = 3$	0.4701	1.2845	1.4399	1.4407	176.48
	(0.1941)	(0.0307)	(0.0238)	(0.0259)	
$q = 4$	0.4176	1.2838	1.4348	1.4214	50.769
	(0.1873)	(0.0309)	(0.0239)	(0.0253)	
$q = 5$	0.4143	1.2836	1.4327	1.4180	43.710
	(0.1798)	(0.0309)	(0.0239)	(0.0244)	
$q = 6$	0.4144	1.2836	1.4315	1.4165	-
	(0.1798)	(0.0310)	(0.0239)	(0.0244)	

Notes: Point estimates based on q-ordered orthogonalized estimators, cross-fitted estimates (100 splits). Robust parametric bootstrap standard errors in parentheses (200 replications). In the last column we report the diagnostic for the orthogonality order q (see Appendix C, bootstrapped with 200 replications), which is asymptotically $\chi^2(4)$ if bias at order q is statistically negligible.

6.2 Empirical estimates

Table 1 shows the estimates of β_0, γ_0, $\sigma_0^2(2)$, and $\sigma_0^2(1)$ for various estimators. These are the plug-in estimator based on the preliminary estimates $\widehat{\eta}_i$ and six estimators based on Neyman-orthogonalized moments, for $1 \leq q \leq 6$. In addition to point estimates, we report robust standard errors based on the parametric bootstrap.[7]

Starting with the substitution parameter γ, the uncorrected estimate is 0.13, which is close to the Cobb-Douglas case. The value of the first-order Neyman-orthogonalized estimate is quite different. However, since the preliminary estimates of the author fixed effects are based on very few observations, we do not expect this estimator to adequately correct for bias. This is confirmed by the fact that all other Neyman-orthogonalized estimates, for $q \in \{2, \ldots, 6\}$, range between 0.41 and 0.77, which is higher than the plug-in estimate, and very different from the first-order orthogonal-

[7]Bootstrap replications are based on Neyman-orthogonalized estimates of β_0, γ_0, $\sigma_0^2(2)$, and $\sigma_0^2(1)$ to order $q = 6$, together with the sample-split estimates $\widehat{\eta}_i$ of author effects. Within each bootstrap replication, we cross-fit the estimates 10 times. Results are based on 200 bootstrap replications. We report robust estimates of bootstrap standard deviations, obtained as interdecile ranges of bootstrap draws divided by $2\Phi^{-1}(0.90) = 2.5631$ (Machado and Parente, 2005). While using robust estimates has little impact on standard errors of parameter estimates in Table 1, this deals with the presence of a few influential bootstrap draws in the calculation of re-allocation effects in Table 2.

ized estimate. Relative to the plug-in, the orthogonalized estimates with $q \geq 2$ all indicate somewhat less complementarity between authors in team production. Notice the stability of estimates for larger values of q. A substitution parameter $\gamma = 0.4$ corresponds to a production function that is supermodular in worker inputs; see Figure 2 in Appendix G for a graphical illustration.

Turning to the other parameters, the estimates of the team size parameter β are virtually unaffected by the orthogonalization. This suggests the bias is limited for this parameter. Its value is close to 1.3, implying that producing a paper with a co-author increases the paper's quality to some extent. Next, the log-error variance $\sigma^2(2)$ in teams of two co-authors is larger when using plug-in estimates (1.6) than when using orthogonalization with $q \geq 2$ (1.4), suggesting that the plug-in and first-order corrected estimates are biased upward. Lastly, the variance $\sigma^2(1)$ in teams of a single author is also larger under the plug-in estimator.

An interesting feature of Table 1 is that point-estimates and standard errors appear to converge as q increases. This suggests the absence of a sharp bias-variance trade-off as a function of q. As we have seen in the case of the Neyman-Scott example in Subsection 5.1, this phenomenon is specific to the model and target parameter of interest, and the variance may converge or diverge as q grows depending on the context.

One potential explanation for convergence in the present setting is that Model (5.3) implies some non-trivial restrictions on the parameters $\gamma, \beta, \sigma^2(1), \sigma^2(2)$ that do not depend on the author-specific effects η_i, as we show in Appendix F. Moment functions that are independent of η are common in panel data settings, obtained e.g. by differencing, quasi-differencing, or functional differencing. In some models, such as models with discrete outcomes, such functions may not exist. In Appendix F we exploit two types of restrictions as robustness checks. Our findings suggest that, while those restrictions seem broadly consistent with the higher-order orthogonal estimates reported in Table 1, using them directly for estimation may lead to very imprecise estimates. In contrast, Table 1 suggests that higher-order Neyman-orthogonality may be a successful approach to approximate such functions in settings where those exist.

In the last column of Table 1 we report our diagnostic statistic for the orthogonality order q, described in Appendix C.[8] A value larger than the 95-th quantile of

[8]The statistic in Appendix C depends on a variance matrix estimate $\widehat{V}$, which we compute based on the parametric bootstrap, cross-fitted 10 times (200 replications).

Table 2: Empirical estimates: average output

	Average output	Counterfactual
Plug-in	8.4434	7.2045
	(0.2038)	(0.1676)
$q = 1$	6.8298	5.4310
	(0.5122)	(0.4260)
$q = 2$	9.0894	8.6759
	(0.5482)	(0.7684)
$q = 3$	6.1291	5.5553
	(0.3778)	(0.3692)
$q = 4$	7.4412	6.6691
	(0.3931)	(0.4407)
$q = 5$	6.9758	6.2593
	(0.2380)	(0.2985)
$q = 6$	7.1168	6.3805
	(0.2614)	(0.3332)

Notes: Average output (the value in the data among co-authored articles is 6.9995*, robust standard error* 0.2231*), and counterfactual average output in a random allocation. Point estimates based on orthogonalized estimators to order q, cross-fitted estimates (100 splits). Robust parametric bootstrap standard errors in parentheses (200 replications).*

the $\chi^2(4)$ distribution (9.488) should be interpreted as evidence that the order-q and order-$(q+1)$ orthogonal moments differ. The values of the statistic reported in the table suggest that biases remain statistically detectable even at $q = 5$. This is likely due to the large sample size, since point estimates are economically similar between $q = 4$ and $q = 6$.

Lastly, we report estimates of average journal quality in a counterfactual scenario where authors are randomly assigned across teams of two co-authors, see (6.1). The first column in Table 2 shows estimates of the average output in the empirical allocation. This quantity can be estimated without bias as the sample mean of the journal quality variable among co-authored articles, which is equal to 7.0. We see that the plug-in estimate is 8.4, larger than the empirical value. In comparison, Neyman-orthogonalized estimates for $q \geq 3$ range between 6.1 and 7.4, and estimates for $q = 5$ and $q = 6$ are closest to the empirical value. The second column in Table 2 shows estimates of average article quality under random assignment of authors to teams, using the plug-in method and Neyman-orthogonalized estimates to order $q \geq 1$.[9] The estimates vary with the order of orthogonalization. When taking $q \geq 4$, estimates range between 6.2 and 6.7. In addition, comparing the two columns of Table 2 shows

[9]To speed up computation, we approximate (6.1) using a random subset of 5000 pairs of authors, for each random sample split (and each bootstrap replication).

that, irrespective of the order of orthogonalization, the estimates of average output are lower in the counterfactual scenario where workers are randomly allocated across teams.

The main takeaway from Table 2 is that randomly allocating authors among teams would tend to lower average paper quality. This is due to two economic forces. The first one is complementarity in production, as reflected by estimates of γ lower than 1. The second force is positive sorting. Indeed, the preliminary estimates of worker fixed effects are positively correlated within teams in the data. In the presence of complementarity, decreasing assortative matching leads to lower output, which is what we find in Table 2.

7 Asymptotic properties

In this section, we show that, under higher-order orthogonality, the estimators $\widehat{\theta}$ and $\widehat{\mu}$ introduced in Subsection 3.2 are $\sqrt{n}$-consistent and asymptotically normal under appropriate assumptions, even if the convergence rate of $\widehat{\eta}_i$ is slower than $\sqrt{n}$. We focus on deriving the asymptotic distribution of $\widehat{\mu}$, assuming that we have already worked out the corresponding asymptotic result of $\widehat{\theta}$. However, the corresponding theory for $\widehat{\theta}$ is actually a special case of our results for $\widehat{\mu}$, where θ is dropped from the arguments, μ is replaced by θ, and u is replaced by $\widetilde{u}$. Thus, our focus on $\widehat{\mu}$ is without loss of generality.

7.1 Notation

For the presentation of the asymptotic theory, it is useful to be explicit about which parameters depend on the sample size and which ones do not. Recall that n is the total number of observations in $(Z_1, ..., Z_N)$, where each Z_i comprises n_i observations. In the asymptotic sequence, we let N and n_i depend on n, although we do not explicitly indicate this dependence. For example, in a panel data model, our assumptions allow both N and T to grow as the number NT of observations tends to infinity.

To indicate the dependence on the sample size, we will write η_n and μ_n instead of η and μ in this section. While the dimension of μ is not changing with n, the true parameter $\mu_{0,n}$ is implicitly defined as the solution of $\sum_{i=1}^{N} \mathbb{E}_{\theta_0,\eta_{0,n}} \left(u(Z_i; \theta_0, \eta_{0,n,i}, \mu)\right) = 0$, which may depend on n. By contrast, the parameter θ and its true value θ_0 are

independent of n.

Remember also that $n = \sum_{i=1}^{N} n_i$, and note that if the observations within each unit i are independent, then we have $\ell(y_i \,|\, x_i; \theta, \eta_{n,i}) = \prod_{j=1}^{n_i} \ell(y_{ij} \,|\, x_{ij}; \theta, \eta_{n,i})$. Hence, in the case of the score for θ,

$$u(Z_i; \theta, \eta_{n,i}) = \sum_{j=1}^{n_i} \frac{\partial \log \ell(y_{ij} \,|\, x_{ij}; \theta, \eta_{n,i})}{\partial \theta}.$$

More generally, whenever $n_i \to \infty$ we expect that u scales linearly with n_i, explaining the scaling of $u(Z_i; \theta, \eta_{n,i})$ and of various other terms in Assumption 1 below.

Finally, in what follows, $D^m_{\eta_{n,i}}$ denotes the derivative operator with respect to $\eta_{n,i}$, and $\mathbb{K}_{q,n,i} = \{m \in \mathbb{N}^{\dim(\eta_{n,i})} : 1 \leq \sum_r m_r \leq q\}$.

7.2 A useful lemma

With this notation in hand, we now state our first assumption.

Assumption 1.

(i) *We have* $\left[\frac{1}{n}\sum_{i=1}^{N} \frac{\partial u^\top(Z_i; \widehat{\theta}, \widehat{\eta}_{n,i}, \widehat{\mu}_n)}{\partial \mu}\right] W \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{N} u(Z_i; \widehat{\theta}, \widehat{\eta}_{n,i}, \widehat{\mu}_n)\right] = o_P(1)$, *for some non-random symmetric positive definite weight matrix* W.

(ii) *As* $n \to \infty$, $(\widehat{\theta}, \widehat{\eta}_n, \widehat{\mu}_n)$ *is contained in a convex neighborhood* $\mathcal{B}_n$ *of* $(\theta_0, \eta_{0,n}, \mu_{0,n})$. *Let* $\mathcal{B}_{n,i}$ *be the convex neighborhood of* $(\theta_0, \eta_{0,n,i}, \mu_{0,n})$ *obtained by intersecting* $\mathcal{B}_n$ *with the parameter subspace for observation* i.

(iii) $\max_i \dim(\eta_{n,i}) = O(1)$.

(iv) *For every* i, *the function* $u(Z_i, \theta, \eta_{n,i}, \mu)$ *is* $(q+1)$ *times continuously differentiable in the parameters* $(\theta, \eta_{n,i}, \mu)$, *and we assume that for all its components all the partial derivatives of* $u(Z_i; \theta, \eta_{n,i}, \mu)$ *up to order* $(q+1)$ *are bounded in absolute value by* $n_i C_{n,i}(Z_i) \geq 0$, *uniformly in the neighborhood* $\mathcal{B}_{n,i}$, *such that* $\frac{1}{n}\sum_{i=1}^{N} n_i \mathbb{E}\left[C_{n,i}(Z_i)^2\right] = O(1)$. *In addition, for all* $m \in \mathbb{K}_{q,n,i}$ *we have* $\mathbb{E}\left[\|D^m_{\eta_{n,i}} u(Z_i; \theta_0, \eta_{0,n,i}, \mu_{0,n})\|^2\right] \leq n_i \widetilde{C}_{n,i}$ *for non-random* $\widetilde{C}_{n,i} \geq 0$ *such that* $\frac{1}{n}\sum_{i=1}^{N} n_i \widetilde{C}^2_{n,i} = O(1)$.

(v) $\widehat{\mu}_n - \mu_{0,n} = o_P(1)$ *and* $\frac{1}{n}\sum_{i=1}^{N} n_i \mathbb{E}\left(\|\widehat{\eta}_{n,i} - \eta_{0,n,i}\|^{2(q+1)}\right) = o(n^{-1})$.

(vi) $\widehat{\theta} = \theta_0 + \frac{1}{n}\sum_{i=1}^{N} \psi_{n,i} + o_P(n^{-1/2})$, *where* $\psi_{n,i}$ *is a function of* Z_i *with* $\mathbb{E}(\psi_{n,i}) = 0$ *and* $\frac{1}{n}\sum_{i=1}^{N} \mathbb{E}\left(\|\psi_{n,i}\|^2\right) = O(1)$.

(vii) The probability limits

$$G_\mu = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{N} \frac{\partial u(Z_i;\theta_0,\eta_{0,n,i},\mu_{0,n})}{\partial \mu^\top}, \quad G_\theta = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{N} \frac{\partial u(Z_i;\theta_0,\eta_{0,n,i},\mu_{0,n})}{\partial \theta^\top}$$

exist, and $\operatorname{rank}(G_\mu) = \dim(\mu)$.

Part (i) in Assumption 1 is satisfied if $\widehat{\mu}_n$ is computed using GMM, see (3.3). In Part (ii), the neighborhood $\mathcal{B}_n$ depends on the sample size n, partly because the number of nuisance parameters of $\eta_{n,i}$ generally depends on n. Part (iii) assumes that the maximal dimension of $\eta_{n,i}$ is bounded as $n \to \infty$. Part (iv) requires the derivatives of the moment functions (properly rescaled) to be suitably bounded. Its second requirement bounds the second moment of those derivatives by n_i, rather than by n_i^2 as the first bound would deliver. Under Assumption 2(i) these derivatives have mean zero, so that their second moment is a variance; when $u(Z_i;\cdot)$ is a sum over the n_i observations in Z_i, as in our panel data and network examples, this variance is of order n_i. The first half of Part (v) is a high-level consistency assumption for $\widehat{\mu}_n$, which can be justified by guaranteeing that the objective function in (3.3) converges uniformly to a population counterpart that has a unique minimum at μ_0. The second half of Part (v) is the rate requirement on the preliminary estimates $\widehat{\eta}_{n,i}$, imposing a rate faster than $n^{-1/2(q+1)}$. Part (vi) requires $\widehat{\theta}$ to be asymptotically linear, in particular requiring $\widehat{\theta} - \theta_0 = O_P(n^{-1/2})$. In the case where $\mu_{0,n} = \theta_0$ this condition is not needed. Lastly, Part (vii) assumes existence of Jacobian matrices and a rank condition.

Lemma 1. *Under Assumption 1 we have*

$$\begin{aligned}
&\sqrt{n}\,(\widehat{\mu}_n - \mu_{0,n}) \\
&\quad = -\left(G_\mu^\top W G_\mu\right)^{-1} G_\mu^\top W \left\{ \frac{1}{\sqrt{n}}\sum_{i=1}^{N} \left[u(Z_i;\theta_0,\eta_{0,n,i},\mu_{0,n}) + G_\theta\,\psi_{n,i} \right] + R_n \right\} + o_P(1),
\end{aligned}$$

where

$$R_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{N} \sum_{m\in\mathbb{K}_{q,n,i}} \frac{1}{m!} \left[D^m_{\eta_{n,i}} u(Z_i;\theta_0,\eta_{0,n,i},\mu_{0,n}) \right] (\widehat{\eta}_{n,i} - \eta_{0,n,i})^m .$$

7.3 Main result

We are now in position to establish the main result of this section, which concerns root-n consistency and asymptotic normality of estimators based on orthogonal equations. For this, we first state our second assumption.

Assumption 2.

(i) *The moment function $u(Z_i;\theta,\eta_{n,i},\mu)$ is Neyman-orthogonal to order q, and furthermore $\sum_{i=1}^N \mathbb{E}\left(u(Z_i;\theta_0,\eta_{0,n,i},\mu_{0,n})\right)=0$.*

(ii) *$\widehat{\eta}_{n,i}$ are independent of $(Z_1,\ldots,Z_N)$ for all i.*

(iii) *The $Z_1,\ldots,Z_N$ are independent across i.*

(iv) *$\xi_{n,i} = u(Z_i;\theta_0,\eta_{0,n,i},\mu_{0,n}) + G_\theta\,\psi_{n,i}$ satisfies Lindeberg's condition,[10] and the following probability limit exists:*

$$V_\xi = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^N \operatorname{Var}\left(\xi_{n,i}\right).$$

Part (i) in Assumption 2 requires u to be Neyman-orthogonal in the sense of Definition 1. Part (ii) requires the preliminary estimates to be independent from the estimation sample. With independent observations, this can be achieved by sample splitting. Part (iii) imposes independence between the Z_i's. We impose this assumption to simplify the presentation. It is straightforward to modify the variance expression in Theorem 2 below to account for particular forms of dependence (e.g., clustered) by using an appropriate expression for the matrix V_ξ introduced in Part (iv).

The following theorem provides an asymptotic characterization of $\widehat{\mu}_n$.

Theorem 2. *Let Assumptions 1 and 2 hold with the same value of $q\in\{1,2,3,\ldots\}$. Then we have*

$$\sqrt{n}\left(\widehat{\mu}_n-\mu_{0,n}\right)\xrightarrow{d}\mathcal{N}\Big(0,\ \Big(G_\mu^\top W\,G_\mu\Big)^{-1}G_\mu^\top W\,V_\xi\,W G_\mu\Big(G_\mu^\top W\,G_\mu\Big)^{-1}\Big).$$

[10] Formally (for scalar $\xi_{n,i}$): for any $\epsilon>0$, $\frac{1}{s_n^2}\sum_{i=1}^N \mathbb{E}\left[\xi_{n,i}^2\cdot\mathbb{1}(|\xi_{n,i}|>\epsilon s_n)\right]\to 0$ as $n\to\infty$, where $s_n^2=\sum_{i=1}^N \operatorname{Var}(\xi_{n,i})$ and $\mathbb{1}$ is the indicator function. Whenever $\mathbb{E}(\xi_{n,i})\neq 0$, the condition is imposed on the centered variables $\xi_{n,i}-\mathbb{E}(\xi_{n,i})$; note that $\sum_{i=1}^N \mathbb{E}(\xi_{n,i})=0$. When $\xi_{n,i}$ is a vector, we impose this condition on $\lambda^\top\xi_{n,i}$ for every non-zero conformable vector λ, and appeal to the Cramér-Wold device.

Note that, although we leave the dependence on q implicit in Theorem 2, the asymptotic variance does depend on the order of orthogonality q that the moment function satisfies. Note also that $\widehat{\mu}_n$ in the theorem is based on a single set of preliminary estimates $\widehat{\eta}_i$. The variability caused by the use of a single sample split can be mitigated through the use of cross-fitting, as we do in the application.

8 Final remarks

In this paper we show how to construct higher-order Neyman-orthogonal moment functions in conditional-likelihood models. We use these functions, together with sample splitting, to reduce bias in estimation. Our application suggests that our higher-order corrections can be effective in network settings with fixed effects. An area of application is to double/debiased machine learning with fixed effects, where the nuisance parameters contain some components, such as low-dimensional functions, for which first-order orthogonality may suffice. However, for such applications it is important to extend the approach to non-likelihood models. As q increases, orthogonalization imposes growing demands on the likelihood structure – to achieve first-order orthogonality it is sufficient for the score to have mean zero, while to achieve second-order orthogonality our approach requires the information identity to hold, for example. This reflects a trade-off between the robustness to nuisance parameters that our method achieves and robustness to model misspecification. In follow-up work (Bonhomme, Jochmans, Newey and Weidner, 2026), we propose a construction of orthogonal functions in semi-parametric models defined by moment conditions. In the absence of a conditional-likelihood structure, our approach exploits independence between observations to achieve higher-order orthogonality.

References

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

Ahmadpoor, M. and B. F. Jones (2019). Decoding team and individual impact in science and invention. *Proceedings of the National Academy of Sciences 116*, 13885–13890.

Angrist, J. D. and B. Frandsen (2022). Machine labor. *Journal of Labor Economics 40*, S97–S140.

Arellano, M. (2003). Discrete choices with panel data. *Investigaciones Economicas XXVII*, 423–458.

Arellano, M. and J. Hahn (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances In Economics and Econometrics*, Volume III. Econometric Society: Cambridge University Press.

Bhattacharyya, A. (1946). On some analogues of the amount of information and their use in statistical estimation. *Sankhyā 8*, 1–14.

Bickel, P. (1982). On adaptive estimation. *Annals of Statistics 10*, 647–671.

Bonhomme, S. (2021). Teams: Heterogeneity, sorting, and complementarity. Mimeo.

Bonhomme, S., K. Jochmans, W. K. Newey, and M. Weidner (2026). Higher-order neyman orthogonality in moment-condition models. *arXiv preprint arXiv:2605.10842*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

Chesher, A. (1991). The effect of measurement error. *Biometrika 78*, 451–462.

Constantine, G. and T. Savits (1996). A multivariate Faa di Bruno formula with applications. *Transactions of the American Mathematical Society 348*, 503–520.

Dhaene, G. and K. Jochmans (2015a). Profile-score adjustments for incidental-parameter problems. Mimeo.

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

Ductor, L., M. Fafchamps, S. Goyal, and M. J. Van der Leij (2014). Social networks and research output. *Review of Economics and Statistics 96*, 936–948.

Evdokimov, K. S. and A. Zeleneev (2023). Simple estimation of semiparametric models with measurement errors. *arXiv preprint arXiv:2306.14311*.

Graham, B. S. (2020). Sparse network asymptotics for logistic regression. *arXiv preprint arXiv:2010.04703*.

Graham, B. S., G. W. Imbens, and G. Ridder (2014). Complementarity and aggregate implications of assortative matching: A nonparametric analysis. *Quantitative Economics 5*, 29–66.

Hahn, J. and J. Hausman (2021). Problems with the control variable approach in achieving unbiased estimates in nonlinear models in the presence of many instruments. *Journal of Quantitative Economics 19*, 39–58.

Hahn, J. and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory 27*, 1152–1191.

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

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacher related policies. *Annual Review of Economics 6*, 801–825.

Jochmans, K. and M. Weidner (2019). Fixed-effect regressions on network data. *Econometrica 87*, 1543–1560.

Kingma, D. P. and M. Welling (2014). Auto-encoding variational bayes. *stat 1050*, 1.

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

Kodrzycki, Y. K. and P. Yu (2006). New approaches to ranking economics journals. *The BE Journal of Economic Analysis & Policy 5*.

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies 69*, 647–666.

Li, H., B. Lindsay, and R. Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B 65*, 191–208.

Machado, J. A. and P. Parente (2005). Bootstrap estimation of covariance matrices via the percentile method. *The Econometrics Journal 8*(1), 70–78.

Mackey, L., V. Syrgkanis, and I. Zadik (2018). Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pp. 3375–3383. PMLR.

McLeish, D. L. and C. G. Small (1994). *Hilbert Space Methods in Probability and*

Statistical Inference. Wiley NY.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*, 1349–1382.

Newey, W. K. and J. M. Robins (2017). Cross-fitting and fast remainder rates for semiparametric estimation. Mimeo.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. In U. Grenander (Ed.), *Probability and Statistics*, pp. 416–444. Wiley NY.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Robins, J., L. Li, E. Tchetgen Tchetgen, and A. van der Vaart (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Annals of Statistics 14*, 1139–1151.

Semenova, V., M. Goldman, V. Chernozhukov, and M. Taddy (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics 14*, 471–510.

Small, C. G. and D. L. McLeish (1988). Generalizations of ancillarity, completeness, and sufficiency in an inference function space. *Annals of Statistics 16*, 534–551.

Small, C. G. and D. L. McLeish (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika 76*, 693–703.

van der Vaart, A. (2014). Higher order tangent spaces and influence functions. *Statistical Science 29*, 679–686.

Waterman, R. P. and B. G. Lindsay (1996). Projected score methods for approximating conditional scores. *Biometrika 83*, 1–13.

Woutersen, T. (2002). Robustness against incidental parameters. Technical report, Research Report.

Wüthrich, K. and Y. Zhu (2021). Omitted variable bias of Lasso-based inference methods: A finite sample analysis. Forthcoming in *Review of Economics and Statistics*.

APPENDIX

A Proofs

A.1 Proof of Theorem 1

Before proving the theorem, it is useful to establish the following lemma, which we prove in Appendix B.

Lemma 2. *Let $q \in \{1, 2, 3, \ldots\}$, and let x be some realization of the covariates. Remember that ∇_η^q and $w_q(z;\theta,\eta)$ are vectors of dimension $k_q = \sum_{p=1}^q d_p$, and that $\Sigma_{w_q w_q}(x;\theta,\eta)$ is a $k_q \times k_q$ matrix. We assume that $\Sigma_{w_q w_q}(x;\theta,\eta)$ is invertible, and we define*

$$\widetilde{w}_q(z;\theta,\eta) = \Sigma_{w_q w_q}(x;\theta,\eta)^{-1} w_q(z;\theta,\eta).$$

Then,

$$\mathbb{E}_{\theta,\eta}\left[(\nabla_\eta^q)^\top \widetilde{w}_q(Z;\theta,\eta)\middle| X = x\right] = \begin{pmatrix} -\mathbb{I}_{d_1} & 0 & 0 & \cdots & 0 \\ 0 & +\mathbb{I}_{d_2} & 0 & \cdots & 0 \\ 0 & 0 & -\mathbb{I}_{d_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (-1)^q\, \mathbb{I}_{d_q} \end{pmatrix}, \quad \text{(A.1)}$$

where the diagonal $k_q \times k_q$ matrix on the right hand side is obtained by stacking $(-1)^p\, \mathbb{I}_{d_p}$ on the diagonal for $p = 1, \ldots, q$, and $\mathbb{I}_{d_p}$ is the identity matrix of dimensions d_p.

Proof of Theorem 1. Define $c_q(x;\theta,\eta,\mu) = \left[\Sigma_{w_q u}(x;\theta,\eta,\mu) - b_q(x;\theta,\eta,\mu)\right]^\top$. In the proof of Lemma 2 in Appendix B we introduce the notation $\widetilde{w}_q^m(z;\theta,\eta)$, $m \in \mathcal{C}_q$, for the elements of the k_q-vector $\widetilde{w}_q(z;\theta,\eta)$. Analogously, we now use $c_q^m(x;\theta,\eta,\mu)$ to denote the columns of the $(\dim u) \times k_q$-matrix $c_q(x;\theta,\eta,\mu)$, that is, $c_q^m(x;\theta,\eta,\mu)$ is a

$(\dim u)$-vector for every $m \in \mathcal{C}_q$. We have

$$\begin{aligned} c_q^m(x;\theta,\eta,\mu) &= \mathbb{E}_{\theta,\eta}\left[\frac{D_\eta^m \ell(y \,|\, x;\theta,\eta)}{\ell(y \,|\, x;\theta,\eta)} u(Z;\theta,\eta,\mu) \,\middle|\, X = x\right] - D_\eta^m \,\mathbb{E}_{\theta,\eta}(u(Z;\theta,\eta,\mu)|X=x) \\ &= \int [D_\eta^m \ell(y \,|\, x;\theta,\eta)]\, u(z;\theta,\eta,\mu) dy - D_\eta^m \int \ell(y \,|\, x;\theta,\eta)\, u(z;\theta,\eta,\mu) dy \\ &= -\sum_{S \subsetneq \{1,\ldots,|m|\}} \int [D_\eta^{m_S}\, \ell(y \,|\, x;\theta,\eta)][D_\eta^{m-S}\, u(z;\theta,\eta,\mu)] dy, \end{aligned} \tag{A.2}$$

where $z = (y, x)$ and in the last step we applied the product rule for differentiation as in (B.3) below, but the term for $S = \{1, \ldots, |m|\}$ cancels with the term that stems from $\Sigma_{w_q u}(x;\theta,\eta,\mu)$, which explains why we only sum over subsets S that are different from $\{1, \ldots, |m|\}$.

Next, by the definition of u_q^* and $A(x;\theta,\eta,\mu)$, we have

$$\begin{aligned} u_q^*(z;\theta,\eta,\mu) &= u(z;\theta,\eta,\mu) - A(x;\theta,\eta,\mu)^\top w_q(z;\theta,\eta) \\ &= u(z;\theta,\eta,\mu) - \left[\Sigma_{w_q u}(x;\theta,\eta,\mu) - b_q(x;\theta,\eta,\mu)\right]^\top \widetilde{w}_q(z;\theta,\eta) \\ &= u(z;\theta,\eta,\mu) - \sum_{v \in \mathcal{C}_q} c_q^v(x;\theta,\eta,\mu)\, \widetilde{w}_q^v(z;\theta,\eta). \end{aligned}$$

Let $m \in \mathcal{C}_q$. Applying the operator D_η^m to the last equation and again using the product rule for differentiation in the same way as before, we find

$$D_\eta^m u_q^*(z;\theta,\eta,\mu) = D_\eta^m u(z;\theta,\eta,\mu) - \sum_{S \subseteq \{1,\ldots,|m|\}} \sum_{v \in \mathcal{C}_q} \left[D_\eta^{m-S}\, c_q^v(x;\theta,\eta,\mu)\right] \left[D_\eta^{m_S}\, \widetilde{w}_q^v(z;\theta,\eta)\right].$$

Applying the conditional expectation operator to this and using Lemma 2 we obtain

$$\begin{aligned} &\mathbb{E}_{\theta,\eta}\left[D_\eta^m u_q^*(Z;\theta,\eta,\mu)\middle| X = x\right] \\ &= \mathbb{E}_{\theta,\eta}\left[D_\eta^m u(Z;\theta,\eta,\mu)\middle| X = x\right] \\ &\quad - \sum_{S \subseteq \{1,\ldots,|m|\}} \sum_{v \in \mathcal{C}_q} \left[D_\eta^{m-S}\, c_q^v(x;\theta,\eta,\mu)\right] \underbrace{\mathbb{E}_{\theta,\eta}\left[D_\eta^{m_S}\, \widetilde{w}_q^v(Z;\theta,\eta)\middle| X = x\right]}_{=(-1)^{|S|}\, \mathbb{1}\{m_S = v\}} \\ &= \mathbb{E}_{\theta,\eta}\left[D_\eta^m u(Z;\theta,\eta,\mu)\middle| X = x\right] - \sum_{\emptyset \neq S \subseteq \{1,\ldots,|m|\}} (-1)^{|S|} \left[D_\eta^{m-S}\, c_q^{m_S}(x;\theta,\eta,\mu)\right], \end{aligned}$$

where in the last step we used that for $S = \emptyset$ the indicator $\mathbb{1}\{m_S = v\}$ is always zero (because $v \in \mathcal{C}_q$ never has length zero), but for $S \neq \emptyset$ there is always exactly one

$v \in \mathcal{C}_q$ that satisfies $m_S = v$, that is, in that second case we just remove the sum over v and replace v by m_S throughout. Next, plugging in the expression for $c_q^m(x;\theta,\eta,\mu)$ in equation (A.2) above we find

$$\begin{aligned}
&\mathbb{E}_{\theta,\eta}\left[D_\eta^m u_q^*(Z;\theta,\eta,\mu)\middle|X=x\right] \\
&= \mathbb{E}_{\theta,\eta}\left[D_\eta^m u(Z;\theta,\eta,\mu)\middle|X=x\right] \\
&\quad + \sum_{\emptyset\neq S\subseteq\{1,\ldots,|m|\}} (-1)^{|S|} D_\eta^{m-S}\left(\sum_{T\subsetneq S}\int [D_\eta^{m_T}\,\ell(y\,|\,x;\theta,\eta)][D_\eta^{m_{S\setminus T}}\,u(z;\theta,\eta,\mu)]dy\right).
\end{aligned}$$

By again using the product rule for differentiation to apply D_η^{m-S} to the product in the last term, we obtain

$$\begin{aligned}
&\mathbb{E}_{\theta,\eta}\left[D_\eta^m u_q^*(Z;\theta,\eta,\mu)\middle|X=x\right] \\
&= \int \ell(y\,|\,x;\theta,\eta)\,[D_\eta^m u(z;\theta,\eta,\mu)]dy \\
&\quad + \sum_{\emptyset\neq S\subseteq\{1,\ldots,|m|\}} (-1)^{|S|} \sum_{T\subsetneq S}\sum_{R\subseteq -S}\int [D_\eta^{m_{R\cup T}}\,\ell(y\,|\,x;\theta,\eta)][D_\eta^{m_{(-S\setminus R)\cup(S\setminus T)}}\,u(z;\theta,\eta,\mu)]dy,
\end{aligned} \tag{A.3}$$

where we write $-S$ for the set $\{1,\ldots,|m|\}\setminus S$. All the terms on the right hand side of the last display equation are of the form $\int [D_\eta^{m_A}\,\ell(y\,|\,x;\theta,\eta)][D_\eta^{m-A}\,u(z;\theta,\eta,\mu)]dy$, for some $A\subseteq\{1,\ldots,|m|\}$, and we can therefore write

$$\mathbb{E}_{\theta,\eta}\left[D_\eta^m u_q^*(Z;\theta,\eta,\mu)\middle|X=x\right] = \sum_{A\subseteq\{1,\ldots,|m|\}} \kappa_A \int [D_\eta^{m_A}\,\ell(y\,|\,x;\theta,\eta)][D_\eta^{m-A}\,u(z;\theta,\eta,\mu)]dy, \tag{A.4}$$

with

$$\kappa_A = \mathbb{1}\{A=\emptyset\} + \sum_{\emptyset\neq S\subseteq\{1,\ldots,|m|\}} (-1)^{|S|}\sum_{T\subsetneq S}\sum_{R\subseteq -S}\mathbb{1}\{R\cup T = A\}.$$

Here, the indicator $\mathbb{1}\{A=\emptyset\}$ accounts for the term $\int \ell(y\,|\,x;\theta,\eta)[D_\eta^m u(z;\theta,\eta,\mu)]dy$ in (A.3), while the second term in κ_A counts the contributions from the triple sum. Our goal is to show that $\kappa_A = 0$ for all $A\subseteq\{1,\ldots,|m|\}$. We analyze two cases separately:

- **Case 1:** For $A=\emptyset$, we note that $R\cup T=\emptyset$ implies $R=T=\emptyset$. Thus, the

indicator $\mathbb{1}\{R \cup T = \emptyset\}$ is non-zero only when $R = \emptyset$ and $T = \emptyset$, implying that

$$\kappa_\emptyset = 1 + \sum_{\emptyset \neq S \subseteq \{1,\ldots,|m|\}} (-1)^{|S|} = \sum_{S \subseteq \{1,\ldots,|m|\}} (-1)^{|S|} = 0,$$

where in the second step we used that $1 = (-1)^{|\emptyset|}$ to include that term into the sum over S, and the final step is the alternating sum result that is also used in the proof of Lemma 2.

- **Case 2:** Next, consider $A \neq \emptyset$. For given A and S, we have

$$\sum_{T \subsetneq S} \sum_{R \subseteq -S} \mathbb{1}\{R \cup T = A\} = \begin{cases} 1 & \text{if } S \not\subseteq A, \\ 0 & \text{if } S \subseteq A. \end{cases} \tag{A.5}$$

If $S \not\subseteq A$ (i.e. not $S \subseteq A$), then (A.5) holds because a solution to the conditions $R \cup T = A$, $T \subsetneq S$, $R \subseteq -S$ exists and is uniquely given by $T = A \cap S$ and $R = A \cap (-S)$. Uniqueness of the pair (T, R) implies that $\sum_{T \subsetneq S} \sum_{R \subseteq -S} \mathbb{1}\{R \cup T = A\} = 1$ in that case. However, if $S \subseteq A$ then no solution for the pair (T, R) exists (because $T = A \cap S$ implies $T = S$ in that case, which contradicts the condition $T \subsetneq S$), implying that the expression in (A.5) is indeed zero then. Using (A.5) we now find that

$$\begin{aligned} \kappa_A &= \sum_{\substack{\emptyset \neq S \subseteq \{1,\ldots,|m|\} \\ S \not\subseteq A}} (-1)^{|S|} \\ &= \sum_{\emptyset \neq S \subseteq \{1,\ldots,|m|\}} (-1)^{|S|} - \sum_{\emptyset \neq S \subseteq A} (-1)^{|S|} \\ &= \left[\sum_{S \subseteq \{1,\ldots,|m|\}} (-1)^{|S|} - (-1)^{|\emptyset|} \right] - \left[\sum_{S \subseteq A} (-1)^{|S|} - (-1)^{|\emptyset|} \right] \\ &= [0 - 1] - [0 - 1] = 0. \end{aligned}$$

We have thus shown that $\kappa_A = 0$ for all $A \subseteq \{1, \ldots, |m|\}$. By (A.4) we thus have $\mathbb{E}_{\theta,\eta}\left[D_\eta^m u_q^*(Z; \theta, \eta, \mu) \middle| X = x \right] = 0$, which can also be written as

$$\mathbb{E}_{\theta,\eta}\left[\nabla_\eta^q u_q^*(Z; \theta, \eta, \mu) \,\middle|\, X = x \right] = 0.$$

Plugging in the true parameter values θ_0 and η_0 we thus find

$$\mathbb{E}\left[\nabla_\eta^q\, u_q^*(Z;\theta_0,\eta_0,\mu)\,\middle|\, X=x\right]=0,$$

and by the law of iterated expectations also

$$\mathbb{E}\left[\nabla_\eta^q\, u_q^*(Z;\theta_0,\eta_0,\mu)\right]=0.$$

Remarkably, this result holds for any value of μ.

□

A.2 Proof of Lemma 1

Let $\tau := (\theta,\mu)$. We write $u_{k,i}(\tau,\eta_i)$ for $u_k(Z_i;\theta,\eta_i,\mu)$, the kth component of the $\dim(u)$-vector $u(Z_i;\theta,\eta_i,\mu)$. Furthermore, compared to the statement of the lemma we drop all subscripts n in the following derivations. In particular, for $\mathbb{K}_{q,n,i}$ we simply write $\mathbb{K}_{q,i}$. By a mean-value expansion of $\widehat{\eta}_i$ around η_{i0} we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^N u_{k,i}(Z_i;\widehat{\theta},\widehat{\eta}_i,\widehat{\mu}) &= \frac{1}{n}\sum_{i=1}^N u_{k,i}(\widehat{\tau},\widehat{\eta}_i)\\
&= \frac{1}{n}\sum_{i=1}^N u_{k,i}(\widehat{\tau},\eta_{i0})\\
&\quad+\frac{1}{n}\sum_{i=1}^N\sum_{m\in\mathbb{K}_{q,i}}\frac{1}{m!}\left[D_{\eta_i}^m\, u_{k,i}(\widehat{\tau},\eta_{i0})\right](\widehat{\eta}_i-\eta_{i0})^m\\
&\quad+\frac{1}{n}\sum_{i=1}^N\sum_{m\in\mathbb{K}_{q+1,i}\setminus\mathbb{K}_{q,i}}\frac{1}{m!}\left[D_{\eta_i}^m\, u_{k,i}(\widehat{\tau},\widetilde{\eta}_i)\right](\widehat{\eta}_i-\eta_{i0})^m,
\end{aligned}$$

where $m!=\prod_r(m_r!)$, and $\widetilde{\eta}_i$ is some value between η_{i0} and $\widehat{\eta}_i$. Next, we perform a mean-value expansion in $\widehat{\tau}$ around τ_0 to obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^N u_{k,i}(Z_i;\widehat{\theta},\widehat{\eta}_i,\widehat{\mu}) &= \frac{1}{n}\sum_{i=1}^N u_{k,i}(\tau_0,\eta_{i0})\\
&\quad+\underbrace{\frac{1}{n}\sum_{i=1}^N\left[\frac{\partial}{\partial\tau}u_{k,i}(\tau_0,\eta_{i0})\right]^\top(\widehat{\tau}-\tau_0)}_{=\left[G_\mu(\widehat{\mu}-\mu_0)+G_\theta(\widehat{\theta}-\theta_0)\right]_k+o_P(\left\|\widehat{\tau}-\tau_0\right\|)}
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\frac{1}{2}\left(\widehat{\tau}-\tau_0\right)^\top \left\{\frac{1}{n}\sum_{i=1}^{N}\left[\frac{\partial^2}{\partial\tau\partial\tau^\top}u_{k,i}(\widetilde{\tau},\eta_{i0})\right]\right\}\left(\widehat{\tau}-\tau_0\right)}_{=:B_{1,k}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{N}\sum_{m\in\mathbb{K}_{q,i}}\frac{1}{m!}\left[D_{\eta_i}^m u_{k,i}(\tau_0,\eta_{i0})\right]\left(\widehat{\eta}_i-\eta_{i0}\right)^m}_{=n^{-1/2}R_{n,k},\text{ the }k\text{'th component of }R_n\text{ defined in the lemma.} } \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{N}\sum_{m\in\mathbb{K}_{q,i}}\frac{1}{m!}\left[D_{\eta_i}^m\frac{\partial}{\partial\tau}u_{k,i}(\overline{\tau},\eta_{i0})\right]^\top\left(\widehat{\tau}-\tau_0\right)\left(\widehat{\eta}_i-\eta_{i0}\right)^m}_{=:B_{2,k}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{N}\sum_{m\in\mathbb{K}_{q+1,i}\setminus\mathbb{K}_{q,i}}\frac{1}{m!}\left[D_{\eta_i}^m u_{k,i}(\widehat{\tau},\widetilde{\eta}_i)\right]\left(\widehat{\eta}_i-\eta_{i0}\right)^m}_{=:B_{3,k}},
\end{aligned}
$$

where $\widetilde{\tau}$ and $\overline{\tau}$ are values between $\widehat{\tau}$ and τ_0. Denote the dimensions of the parameters θ and μ by d_θ and d_μ, respectively. Our assumptions guarantee that

$$
\begin{aligned}
|B_{1,k}| &\leq \frac{(d_\theta+d_\mu)^2}{2}\left\|\widehat{\tau}-\tau_0\right\|^2\frac{1}{n}\sum_{i=1}^{N}n_i C(Z_i) \\
&\leq \frac{(d_\theta+d_\mu)^2}{2}\left\|\widehat{\tau}-\tau_0\right\|^2\underbrace{\left(\frac{1}{n}\sum_{i=1}^{N}n_i\left[C(Z_i)\right]^2\right)^{1/2}}_{=O_P(1)}\underbrace{\left(\frac{1}{n}\sum_{i=1}^{N}n_i\right)^{1/2}}_{=O(1)} = O_P\left(\left\|\widehat{\tau}-\tau_0\right\|^2\right),
\end{aligned}
$$

$$
\begin{aligned}
|B_{2,k}| &\leq (d_\theta+d_\mu)\left\|\widehat{\tau}-\tau_0\right\|\frac{1}{n}\sum_{i=1}^{N}n_i C(Z_i)\sum_{m\in\mathbb{K}_{q,i}}\frac{1}{m!}\left|\left(\widehat{\eta}_i-\eta_{i0}\right)^m\right| \\
&\leq (d_\theta+d_\mu)\left\|\widehat{\tau}-\tau_0\right\|\underbrace{\left(\frac{1}{n}\sum_{i=1}^{N}n_i\left[C(Z_i)\right]^2\right)^{1/2}}_{=O_P(1)}\underbrace{\left(\frac{1}{n}\sum_{i=1}^{N}n_i\left[\sum_{m\in\mathbb{K}_{q,i}}\frac{1}{m!}\left|\left(\widehat{\eta}_i-\eta_{i0}\right)^m\right|\right]^2\right)^{1/2}}_{=o_P(1)} \\
&= o_P\left(\left\|\widehat{\tau}-\tau_0\right\|\right),
\end{aligned}
$$

$$|B_{3,k}| \leq \frac{1}{n}\sum_{i=1}^{N} n_i C(Z_i) \left\|\widehat{\eta}_i - \eta_{i0}\right\|^{q+1} \underbrace{\sum_{m \in \mathbb{K}_{q+1,i} \setminus \mathbb{K}_{q,i}} \frac{1}{m!}}_{=O(1)}$$
$$= O(1) \underbrace{\left(\frac{1}{n}\sum_{i=1}^{N} n_i \left[C(Z_i)\right]^2\right)^{1/2}}_{=O_P(1)} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{N} n_i \left\|\widehat{\eta}_i - \eta_{i0}\right\|^{2(q+1)}\right)^{1/2}}_{=o_P(n^{-1/2})} = o_P(n^{-1/2}).$$

Here, in addition to our assumption we also used the Cauchy-Schwarz inequality. We have thus shown that

$$\frac{1}{n}\sum_{i=1}^{N} u(Z_i; \widehat{\theta}, \widehat{\eta}_i, \widehat{\mu}) = G_\mu(\widehat{\mu} - \mu_0) + G_\theta(\widehat{\theta} - \theta_0) + \frac{1}{n}\sum_{i=1}^{N} u(Z_i; \theta_0, \eta_{i0}, \mu_0) + n^{-1/2} R_n$$
$$+ O_P\left(\left\|\widehat{\tau} - \tau_0\right\|^2\right) + o_P\left(\left\|\widehat{\tau} - \tau_0\right\|\right) + o_P(n^{-1/2}).$$

Using our assumptions on the convergence of $\widehat{\mu}$ and $\widehat{\theta}$ we thus have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{N} u(Z_i; \widehat{\theta}, \widehat{\eta}_i, \widehat{\mu}) = G_\mu\left[\sqrt{n}(\widehat{\mu} - \mu_0)\right] + o_P\Big(1 + \left\|\sqrt{n}(\widehat{\mu} - \mu_0)\right\|\Big)$$
$$+ \frac{1}{\sqrt{n}}\sum_{i=1}^{N}\left[u(Z_i; \theta_0, \eta_{i0}, \mu_0) + G_\theta\,\psi_i\right] + R_n + o_P(1). \qquad \text{(A.6)}$$

By Assumption 1(i) we have

$$\left[\frac{1}{n}\sum_{i=1}^{N} \frac{\partial u^\top(Z_i; \widehat{\theta}, \widehat{\eta}_i, \widehat{\mu})}{\partial \mu}\right] W \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{N} u(Z_i; \widehat{\theta}, \widehat{\eta}_i, \widehat{\mu})\right] = o_P(1).$$

By using Assumption 1(iv) and (vii) we thus have $G_\mu^\top W\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{N} u(Z_i; \widehat{\theta}, \widehat{\eta}_i, \widehat{\mu})\right] = o_P(1)$, and plugging the approximation in (A.6) then gives

$$o_P(1) = G_\mu^\top W\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{N} u(Z_i; \widehat{\theta}, \widehat{\eta}_i, \widehat{\mu})\right]$$
$$= \left(G_\mu^\top W G_\mu\right)\left[\sqrt{n}(\widehat{\mu} - \mu_0)\right] + o_P\Big(1 + \left\|\sqrt{n}(\widehat{\mu} - \mu_0)\right\|\Big)$$
$$+ G_\mu^\top W\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{N}\left[u(Z_i; \theta_0, \eta_{i0}, \mu_0) + G_\theta\,\psi_i\right] + R_n\right\} + o_P(1).$$

Since $G_\mu^\top W G_\mu$ is full rank, absorbing the $o_P(\|\sqrt{n}(\widehat{\mu}-\mu_0)\|)$ term into $G_\mu^\top W G_\mu$ and solving for $\sqrt{n}(\widehat{\mu}-\mu_0)$ gives the statement of the lemma.

A.3 Proof of Theorem 2

We again drop all subscripts n in the derivations. Let $\xi_i = u(Z_i;\theta_0,\eta_{i0},\mu_0) + G_\theta \psi_i$. From Lemma 1, we have

$$\sqrt{n}(\widehat{\mu}-\mu_0) = -(G_\mu^\top W G_\mu)^{-1} G_\mu^\top W \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^N \xi_i + R_n \right\} + o_P(1).$$

We will show that $R_n = o_P(1)$ under our assumptions. First, by Assumption 2(i), the moment function is Neyman-orthogonal to order q, which implies

$$\mathbb{E}\left[D_{\eta_i}^m u(Z_i;\theta_0,\eta_{i0},\mu_0) \right] = 0$$

for all $m \in \mathbb{K}_{q,i}$. Therefore, R_n is a sum of mean-zero terms. Next, by Assumption 2(ii), (iii), the summands of R_n are uncorrelated across i (since, conditional on $(\widehat{\eta}_1,\ldots,\widehat{\eta}_N)$, they are independent with zero conditional means), so that $\mathbb{E}[\|R_n\|^2]$ is $\frac{1}{n}\sum_{i=1}^N$ of their second moments. By the second-moment bound in Assumption 1(iv) and Assumption 2(ii), the latter are bounded by a constant times $n_i \widetilde{C}_{n,i}\ c_{n,i}$, where $c_{n,i} := \left(\mathbb{E}\left\|\widehat{\eta}_i - \eta_{i0}\right\|^{2(q+1)}\right)^{\frac{1}{q+1}}$. Here we used that, by Lyapunov's inequality, $\mathbb{E}\left[\left((\widehat{\eta}_i - \eta_{i0})^m\right)^2\right] \le \mathbb{E}\left\|\widehat{\eta}_i - \eta_{i0}\right\|^{2|m|} \le c_{n,i}^{|m|}$, and that $c_{n,i}^{|m|} \le c_{n,i}$ for all $1 \le |m| \le q$ once n is sufficiently large, because Assumption 1(v) together with $n_i \ge 1$ implies $\max_i c_{n,i} = o(1)$. Splitting $n_i = \sqrt{n_i}\sqrt{n_i}$ and applying the Cauchy-Schwarz inequality as above,

$$\frac{1}{n}\sum_{i=1}^N n_i \widetilde{C}_{n,i}\, c_{n,i} \le \underbrace{\left(\frac{1}{n}\sum_{i=1}^N n_i \widetilde{C}_{n,i}^2\right)^{1/2}}_{=O(1)} \underbrace{\left(\frac{1}{n}\sum_{i=1}^N n_i\, c_{n,i}^2\right)^{1/2}}_{=o(n^{-1/(q+1)})},$$

where the second factor follows from Jensen's inequality applied to the concave map $x \mapsto x^{2/(q+1)}$ (note that $c_{n,i}^2 = (c_{n,i}^{q+1})^{2/(q+1)}$) together with $\frac{1}{n}\sum_{i=1}^N n_i = 1$, and Assumption 1(v). Hence $\mathbb{E}[\|R_n\|^2] = o(1)$. By Chebyshev's inequality we thus have

$R_n = o_P(1)$. Thus, we have

$$\sqrt{n}(\widehat{\mu} - \mu_0) = -(G_\mu^\top W G_\mu)^{-1} G_\mu^\top W \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \xi_i \right\} + o_P(1).$$

By Assumption 2(ii), (iii), the terms ξ_i are independent across i. Furthermore, by Assumption 2(iv), they satisfy Lindeberg's condition and have a well-defined variance limit V_ξ. Therefore, by the Lindeberg-Feller Central Limit Theorem:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{N} \xi_i \xrightarrow{d} \mathcal{N}(0, V_\xi).$$

The conclusion follows by the continuous mapping theorem, giving us

$$\sqrt{n}(\widehat{\mu} - \mu_0) \xrightarrow{d} \mathcal{N}\left(0, (G_\mu^\top W G_\mu)^{-1} G_\mu^\top W V_\xi W G_\mu (G_\mu^\top W G_\mu)^{-1}\right).$$

SUPPLEMENT TO:
"A NEYMAN-ORTHOGONALIZATION APPROACH TO
THE INCIDENTAL PARAMETER PROBLEM
IN LIKELIHOOD MODELS"

Stéphane Bonhomme, Koen Jochmmans, and Martin Weidner

B Proof of Lemma 2

By taking derivatives of $\int \ell(y|x;\theta,\eta)dy = 1$ with respect to η, we obtain

$$\int [\nabla_\eta^q \ell(y|x;\theta,\eta)]dy = 0,$$

which can also be written as $\mathbb{E}_{\theta,\eta}[w_q(Z;\theta,\eta)|X=x] = 0$. Since $\Sigma_{w_q w_q}(X;\theta,\eta)$ does not depend on Y we also have $\mathbb{E}_{\theta,\eta}[\widetilde{w}_q(Z;\theta,\eta)|X=x] = 0$. Using this and the definition of $\widetilde{w}_q$ we obtain

$$\begin{aligned}
\mathbb{I}_{k_q} &= \Sigma_{w_q w_q}(x;\theta,\eta)^{-1}\Sigma_{w_q w_q}(x;\theta,\eta) \\
&= \Sigma_{w_q w_q}(x;\theta,\eta)^{-1}\mathbb{E}_{\theta,\eta}[w_q(Z;\theta,\eta)w_q(Z;\theta,\eta)^\top|X=x] \\
&= \mathbb{E}_{\theta,\eta}[\widetilde{w}_q(Z;\theta,\eta)w_q(Z;\theta,\eta)^\top|X=x] \\
&= \int \widetilde{w}_q(z;\theta,\eta)[\nabla_\eta^q \ell(y|x;\theta,\eta)]^\top dy. \qquad \text{(B.1)}
\end{aligned}$$

According to its definition in Subsection 3.1, the elements of the k_q-vector operator ∇_η^q are given by

$$D_\eta^m = \frac{\partial^p}{\partial\eta_{m_1}\cdots\partial\eta_{m_p}},$$

and are uniquely labeled by vectors of integers $m = (m_1,\ldots,m_p)$ of length $p \in \{1,\ldots,q\}$ in the following set

$$\mathcal{C}_q = \bigcup_{p\in\{1,\ldots,q\}} \{m = (m_1,\ldots,m_p) : 1 \le m_1 \le \cdots \le m_p \le d_\eta\}.$$

Analogously, we now introduce the notation $\widetilde{w}_q^m(z;\theta,\eta)$, $m \in \mathcal{C}_q$, to uniquely denote the elements of the vector $\widetilde{w}_q(z;\theta,\eta)$, which is also a vector of length $k_q = |\mathcal{C}_q|$. With

that notation, the result in display (B.1) can equivalently be written as

$$\forall r, m \in \mathcal{C}_q : \qquad \int \widetilde{w}_q^r(z;\theta,\eta)\,[D_\eta^m \ell(y|x;\theta,\eta)]\,dy = \mathbb{1}\{r=m\}. \tag{B.2}$$

Since $\mathbb{E}_{\theta,\eta}[\widetilde{w}_q(Z;\theta,\eta)|X=x]=0$, we also have

$$\int \widetilde{w}_q^r(z;\theta,\eta)\,\ell(y|x;\theta,\eta)\,dy = 0.$$

For the empty vector () of length zero we have $D_\eta^{()}\ell(y|x;\theta,\eta) = \ell(y|x;\theta,\eta)$. Using this notation we can combine the result in the last two displays to find that for all $r \in \mathcal{C}_q$ and all $v \in \mathcal{C}_q \cup \{()\}$ we have

$$\int \widetilde{w}_q^r(z;\theta,\eta)\,[D_\eta^v \ell(y|x;\theta,\eta)]\,dy = \mathbb{1}\{r=v\}.$$

For $k \in \{1,2,3,\ldots\}$, vector $t=(t_1,\ldots,t_k) \in \mathcal{C}_q$, and a subset $S \subseteq \{1,\ldots,k\}$, let t_S denote the vector formed by keeping only the indices in S, and let $t_{-S} = t_{\{1,\ldots,k\}\setminus S}$ be the vector of the remaining elements. Then, by applying D_η^t to the last display and using the product rule for differentiation we obtain

$$\sum_{S\subseteq\{1,\ldots,k\}} \int [D_\eta^{t_S}\widetilde{w}_q^r(z;\theta,\eta)]\,[D_\eta^{t_{-S}} D_\eta^v \ell(y|x;\theta,\eta)]\,dy = 0. \tag{B.3}$$

Of course, we have $D_\eta^t D_\eta^v = D_\eta^{(t,v)}$, and instead of distinguishing between t and v we can also just write m for (t,v) combined. The last display equation then implies that for any nonempty subset $T \subseteq \{1,2,\ldots,|m|\}$ we have (just set $t=m_T$ and $v=m_{-T}$ in the last display result—in doing so, it was important that above we allowed for v to be the empty vector):

$$\sum_{S\subseteq T} \int [D_\eta^{m_S}\widetilde{w}_q^r(z;\theta,\eta)]\,[D_\eta^{m_{-S}} \ell(y|x;\theta,\eta)]\,dy = 0.$$

Now, consider the following linear combination of the result in the last display,

$$\sum_{\substack{T\subseteq\{1,2,\ldots,|m|\}\\ T\neq\emptyset}} (-1)^{|T|} \sum_{S\subseteq T} \int [D_\eta^{m_S}\widetilde{w}_q^r(z;\theta,\eta)]\,[D_\eta^{m_{-S}} \ell(y|x;\theta,\eta)]\,dy = 0. \tag{B.4}$$

For a fixed $S \subseteq \{1, 2, \ldots, |m|\}$, the total coefficient of the term $\int [D_\eta^{m_S} \widetilde{w}_q^r]\, [D_\eta^{m-S} \ell]\, dy$ in this linear combination is given by

$$\sum_{\substack{T\,:\,S \subseteq T \subseteq \{1,\ldots,|m|\} \\ T \neq \emptyset}} (-1)^{|T|} = \begin{cases} -1 & \text{if } S = \emptyset, \\ (-1)^{|m|} & \text{if } S = \{1, \ldots, |m|\}, \\ 0 & \text{otherwise.} \end{cases} \tag{B.5}$$

To see that the result in the last display holds, notice first that for $S = \emptyset$ the sum is simply

$$\sum_{T\,:\,\emptyset \neq T} \subseteq \{1, \ldots, |m|\}(-1)^{|T|} = -1 + \underbrace{\sum_{T \subseteq \{1,\ldots,|m|\}} (-1)^{|T|}}_{=0} = -1,$$

where $\sum_{T \subseteq \{1,\ldots,|m|\}} (-1)^{|T|} = \sum_{r=0}^{|m|} \binom{|m|}{r} (-1)^r = (1-1)^{|m|} = 0$ is a classic alternating sum result, which holds for all $|m| \geq 1$. Next, for $S = \{1, \ldots, |m|\}$, the left hand side of (B.5) only sums over one element, $T = \{1, \ldots, |m|\}$, and we thus get $(-1)^{|T|} = (-1)^{|m|}$ for the sum. Finally, if $S \neq \emptyset$ and $S \neq \{1, \ldots, |m|\}$, then the left hand side of (B.5) can be written as

$$\sum_{T\,:\,S \subseteq T \subseteq \{1,\ldots,|m|\}} (-1)^{|T|} = \sum_{R \subseteq -S} (-1)^{|S|+|R|} = (-1)^{|S|} \underbrace{\sum_{R \subseteq -S} (-1)^{|R|}}_{=0} = 0$$

where we replaced the sum over T by a sum over R such that $T = S \cup R$, with $-S = \{1, \ldots, |m|\} \setminus S$, and in the final step we used the alternating sum result again.

Using (B.5), our linear combination in (B.4) equals

$$-\int [D_\eta^m \widetilde{w}_q^r(z; \theta, \eta)]\, \ell(y|x; \theta, \eta)\, dy + (-1)^{|m|} \int \widetilde{w}_q^r(z; \theta, \eta) [D_\eta^m \ell(y|x; \theta, \eta)]\, dy = 0.$$

Together with (B.2) we thus find

$$\int [D_\eta^m \widetilde{w}_q^r(z; \theta, \eta)]\, \ell(y|x; \theta, \eta)\, dy = (-1)^{|m|}\, \mathbb{1}\{r = m\},$$

which in vector notation can be written as (A.1).

C Order q of orthogonality

C.1 A diagnostic for q

Consider a moment function $u(Z_i;\theta,\eta_i,\mu)$, and denote its q-orthogonal counterparts based on Theorem 1 as $u_q^*(Z_i;\theta,\eta_i,\mu)$, for $q=1,2,....$ Given some order q we propose the following diagnostic:

$$S_q = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \left[u_q^*(Z_i;\widehat{\theta},\widehat{\eta}_i,\widehat{\mu}) - u_{q+1}^*(Z_i;\widehat{\theta},\widehat{\eta}_i,\widehat{\mu}) \right] \right\|^2_{\widehat{V}^{-1}},$$

where $\widehat{\theta}$ and $\widehat{\mu}$ are based on $(q+1)$-orthogonal functions, and $\widehat{V}$ is a consistent estimator of the variance

$$V = \lim_{n\to\infty} \mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{N} (u_q^*(Z_i;\widehat{\theta},\eta_i,\widehat{\mu}) - u_{q+1}^*(Z_i;\widehat{\theta},\eta_i,\widehat{\mu})) \right].$$

As a special case, suppose that one wishes to estimate θ, and $u(Z_i;\theta,\eta_i)$ does not depend on μ. Then, since $\widehat{\theta}$ is based on u_{q+1}^* that is uncorrelated with $u_q^* - u_{q+1}^*$, a consistent estimator of V is

$$\widehat{V} = \frac{1}{n} \sum_{i=1}^{N} (u_q^*(Z_i;\widehat{\theta},\widehat{\eta}_i) - u_{q+1}^*(Z_i;\widehat{\theta},\widehat{\eta}_i))(u_q^*(Z_i;\widehat{\theta},\widehat{\eta}_i) - u_{q+1}^*(Z_i;\widehat{\theta},\widehat{\eta}_i))^\top.$$

More generally, a consistent estimator of V is available using standard Taylor expansions and Lemma 1. We provide an example in the case of the Neyman-Scott model in the next subsection.

Under the conditions of Theorem 2, which in particular guarantee that the estimates $\widehat{\eta}_i$ converge at a fast enough rate, S_q converges in distribution to a χ^2_r random variable, where $r = \dim u$. In practice, we thus suggest comparing S_q to the quantiles of the χ^2_r distribution.

C.2 Illustration in the Neyman-Scott model

Consider the estimation of $\mu = {}^1\!/_N \sum_{i=1}^N h(\eta_i)$ in the Neyman-Scott model introduced in Subsection 2.1. Here we assume that T – the number of periods used in estimation – is fixed as N tends to infinity, and consequently we use $\sqrt{N}$ in order to scale all

cross-sectional average quantities. Suppose first that σ is known. By (5.1) we have

$$\begin{aligned}
&u_q^*(Y_i;\sigma^2,\eta_i,\mu) - u_{q+1}^*(Y_i;\sigma^2,\eta_i,\mu) \\
&= -\sigma^{q+1}T^{-\frac{q+1}{2}}\frac{1}{(q+1)!}\nabla_\eta^{(q+1)}h(\eta_i)H_{q+1}\left(\frac{\sum_{j=1}^T(Y_{ij}-\eta_i)}{\sqrt{T}\sigma}\right),
\end{aligned}$$

and

$$V = \lim_{N\to\infty}\sigma^{2q+2}T^{-q-1}\frac{1}{(q+1)!}\frac{1}{N}\sum_{i=1}^N[\nabla_\eta^{(q+1)}h(\eta_i)]^2.$$

Hence, again treating σ as known, the diagnostic is

$$S_q = \frac{\left[\sum_{i=1}^N \nabla_\eta^{(q+1)}h(\widehat{\eta}_i)H_{q+1}\left(\frac{\sum_{j=1}^T(Y_{ij}-\widehat{\eta}_i)}{\sqrt{T}\sigma}\right)\right]^2}{(q+1)!\sum_{i=1}^N[\nabla_\eta^{(q+1)}h(\widehat{\eta}_i)]^2}. \tag{C.1}$$

With σ unknown, we have, using $\widehat{\sigma}$ in (2.11),

$$\begin{aligned}
&u_q^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu) - u_{q+1}^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu) \\
&= -\widehat{\sigma}^{q+1}T^{-\frac{q+1}{2}}\frac{1}{(q+1)!}\nabla_\eta^{(q+1)}h(\eta_i)H_{q+1}\left(\frac{\sum_{j=1}^T(Y_{ij}-\eta_i)}{\sqrt{T}\widehat{\sigma}}\right) \\
&= -\sigma^{q+1}T^{-\frac{q+1}{2}}\frac{1}{(q+1)!}\nabla_\eta^{(q+1)}h(\eta_i)H_{q+1}\left(\frac{\sum_{j=1}^T(Y_{ij}-\eta_i)}{\sqrt{T}\sigma}\right) \\
&\quad + \frac{1}{2T}\widetilde{\sigma}^{q-1}T^{-\frac{q-1}{2}}\frac{1}{(q-1)!}\nabla_\eta^{(q+1)}h(\eta_i)H_{q-1}\left(\frac{\sum_{j=1}^T(Y_{ij}-\eta_i)}{\sqrt{T}\widetilde{\sigma}}\right)(\widehat{\sigma}^2-\sigma^2),
\end{aligned}$$

for $\widetilde{\sigma}$ between σ and $\widehat{\sigma}$, where we have used the expression for the derivatives of Hermite polynomials.

Hence, for $q \geq 2$ we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(u_q^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu) - u_{q+1}^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu)\right) \\
&= -\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma^{q+1}T^{-\frac{q+1}{2}}\frac{1}{(q+1)!}\nabla_\eta^{(q+1)}h(\eta_i)H_{q+1}\left(\frac{\sum_{j=1}^T(Y_{ij}-\eta_i)}{\sqrt{T}\sigma}\right) + o_P(1),
\end{aligned}$$

and we rely on the diagnostic S_q given by (C.1) even though σ is estimated. Now, for

$q = 1$ we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(u_1^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu) - u_2^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu)\right) \\
&= -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sigma^2 T^{-1}\frac{1}{2}\nabla_\eta^{(2)}h(\eta_i)\left(\left(\frac{\sum_{j=1}^{T}(Y_{ij}-\eta_i)}{\sqrt{T}\sigma}\right)^2 - 1\right) \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{2T}\nabla_\eta^{(2)}h(\eta_i)\left(\frac{1}{N(T-1)}\sum_{i'=1}^{N}\sum_{j=1}^{T}(Y_{i'j}-\overline{Y}_{i'})^2 - \sigma^2\right) + o_P(1),
\end{aligned}$$

where we have used the expressions of H_2 and $\widehat{\sigma}^2$. That is,

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(u_1^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu) - u_2^*(Y_i;\widehat{\sigma}^2,\eta_i,\mu)\right) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{2T}\nabla_\eta^{(2)}h(\eta_i)\left[\frac{1}{N(T-1)}\sum_{i'=1}^{N}\sum_{j=1}^{T}(Y_{i'j}-\overline{Y}_{i'})^2 - \left(\frac{\sum_{j=1}^{T}(Y_{ij}-\eta_i)}{\sqrt{T}}\right)^2\right] + o_P(1).
\end{aligned}$$

In this case,

$$V = \lim_{N\to\infty}\sigma^4 T^{-2}\frac{1}{2}\left[\frac{1}{N}\sum_{i=1}^{N}[\nabla_\eta^{(2)}h(\eta_i)]^2 + \frac{1}{T-1}\left(\frac{1}{N}\sum_{i=1}^{N}\nabla_\eta^{(2)}h(\eta_i)\right)^2\right],$$

and we rely on

$$S_1 = \frac{\left\{\sum_{i=1}^{N}\nabla_\eta^{(2)}h(\widehat{\eta}_i)\left[\left(\frac{\sum_{j=1}^{T}(Y_{ij}-\widehat{\eta}_i)}{\sqrt{T}\widehat{\sigma}}\right)^2 - \frac{1}{N(T-1)}\sum_{i'=1}^{N}\sum_{j=1}^{T}\left(\frac{Y_{i'j}-\overline{Y}_{i'}}{\widehat{\sigma}}\right)^2\right]\right\}^2}{2\left[\sum_{i=1}^{N}[\nabla_\eta^{(2)}h(\widehat{\eta}_i)]^2 + \frac{1}{N(T-1)}\left(\sum_{i=1}^{N}\nabla_\eta^{(2)}h(\widehat{\eta}_i)\right)^2\right]}.$$

Remark 2. (sample splitting) When implementing the diagnostic in this example, we use a small number T of time periods in estimation. Keeping T fixed – while at the same time having $\widehat{\eta}_i \xrightarrow{p} \eta_i$ – is important for the theory in order to guarantee that the impact of the estimation error in $\widehat{\eta}_i$ does not affect the asymptotic distribution of S_q. In practice, one can cross-fit by taking multiple subsets of T time periods (along with different $\widehat{\eta}_i$'s) and average over them to reduce variability.

D Implementation in nonlinear regression

D.1 Expression for M

Following Constantine and Savits (1996), consider a multivariate function

$$h\left(x_1,\ldots,x_d\right)=f\left[g^{(1)}\left(x_1,\ldots,x_d\right),\ldots,g^{(m)}\left(x_1,\ldots,x_d\right)\right],$$

$h_\nu=D_{\mathbf{x}}^{\nu}h\left(\mathbf{x}^0\right),f_{\boldsymbol{\lambda}}=D_{\mathbf{y}}^{\boldsymbol{\lambda}}f\left(\mathbf{y}^0\right),g_{\boldsymbol{\mu}}^{(i)}=D_{\mathbf{x}}^{\mu}g^{(i)}\left(\mathbf{x}^0\right),\mathbf{g}_{\boldsymbol{\mu}}=\left(g_{\boldsymbol{\mu}}^{(1)},\ldots,g_{\boldsymbol{\mu}}^{(m)}\right)$. The Faà di Bruno formula is (Theorem 2.1 in Constantine and Savits, 1996):

$$h_{\boldsymbol{\nu}}=\sum_{1\leq|\boldsymbol{\lambda}|\leq n}f_{\boldsymbol{\lambda}}\underbrace{\sum_{s=1}^{n}\sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})}(\boldsymbol{\nu}!)\prod_{j=1}^{s}\frac{\left[\mathbf{g}_{\boldsymbol{\ell}_j}\right]^{\mathbf{k}_j}}{(\mathbf{k}_j!)\left[\boldsymbol{\ell}_j!\right]^{|\mathbf{k}_j|}}}_{\text{elements of }M},$$

where $n=|\boldsymbol{\nu}|$, and

$$p_s(\boldsymbol{\nu},\boldsymbol{\lambda})=\Bigg\{(\mathbf{k}_1,\ldots,\mathbf{k}_s;\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_s):|\mathbf{k}_i|>0,$$
$$\mathbf{0}\prec\boldsymbol{\ell}_1\prec\cdots\prec\boldsymbol{\ell}_s,\sum_{i=1}^{s}\mathbf{k}_i=\boldsymbol{\lambda}\text{ and }\sum_{i=1}^{s}|\mathbf{k}_i|\,\boldsymbol{\ell}_i=\boldsymbol{\nu}\Bigg\}.$$

D.2 Useful properties of the normal distribution

We first consider the univariate normal case.

Lemma 3. *For $m\in\mathbb{R}$ and $\sigma\in(0,\infty)$, let $Y\sim\mathcal{N}(m,\sigma^2)$, with corresponding likelihood function*

$$\ell(y\,|\,m,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-m)^2}{2\,\sigma^2}\right).$$

Let $j,k\in\{0,1,2,\ldots\}$, and define

$$\kappa_{jk}:=\mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y\,|\,m,\sigma)}\frac{\partial^j\ell(Y\,|\,m,\sigma)}{(\partial m)^j}\frac{1}{\ell(Y\,|\,m,\sigma)}\frac{\partial^k\ell(Y\,|\,m,\sigma)}{(\partial m)^k}\right],$$
$$\rho_j:=\mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y\,|\,m,\sigma)}\frac{\partial^j\ell(Y\,|\,m,\sigma)}{(\partial m)^j}\frac{\partial\log\ell(Y\,|\,m,\sigma)}{\partial\sigma}\right].$$

Then,

$$\kappa_{jk} = \mathbb{1}\Big\{j = k\Big\}\,\frac{j!}{\sigma^{2j}}\,, \qquad\qquad \rho_j = \mathbb{1}\Big\{j = 2\Big\}\frac{2}{\sigma^3}\,.$$

Let $\phi(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-y^2/2\right)$ and $\phi^{(j)}(y) = \frac{d^j\phi(y)}{dy^j}$. Hermite polynomials are defined by $h_j(y) = (-1)^j[\phi(y)]^{-1}\,\phi^{(j)}(y)$. The proof of Lemma 3 is given in Subsection D.4. It crucially relies on the following orthogonality property of Hermite polynomials:

$$\int_{-\infty}^{\infty} h_j(y)\,h_k(y)\,\phi(y)\,dy = \mathbb{1}\Big\{j = k\Big\}\,j!\,. \tag{D.1}$$

The result in Lemma 3 is sufficient for our purposes, but more general results can be derived.[11]

Next, we consider a vector of independent normal variables with heteroscedastic means and variances.

Lemma 4. *Let $d \in \{1, 2, 3, \ldots\}$. For $m \in \mathbb{R}^d$ and $\sigma \in (0,\infty)^d$, let $\Sigma(\sigma)$ be the $d \times d$ diagonal matrix with diagonal entries σ_i^2, and let $Y \sim \mathcal{N}(m, \Sigma(\sigma))$. The corresponding likelihood function reads*

$$\ell(y \,|\, m, \sigma) = \prod_{i=1}^{d} \ell(y_i \,|\, m_i, \sigma_i), \qquad \ell(y_i \,|\, m_i, \sigma_i) = \left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(y_i - m_i)^2}{2\,\sigma_i^2}\right)\right].$$

Let $j, k \in \{0, 1, 2, \ldots\}^d$, $j^* = \sum_{i=1}^d j_i$, $k^* = \sum_{i=1}^d k_i$, and define

$$\kappa(j,k) := \mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y \,|\, m, \sigma)}\,\frac{\partial^{j^*}\ell(Y \,|\, m, \sigma)}{\prod_{i=1}^d (\partial m_i)^{j_i}}\,\frac{1}{\ell(Y \,|\, m, \sigma)}\,\frac{\partial^{k^*}\ell(Y \,|\, m, \sigma)}{\prod_{i=1}^d (\partial m_i)^{k_i}}\right],$$
$$\rho(j,i') := \mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y \,|\, m, \sigma)}\,\frac{\partial^{j^*}\ell(Y \,|\, m, \sigma)}{\prod_{i=1}^d (\partial m_i)^{j_i}}\,\frac{\partial \log \ell(Y \,|\, m, \sigma)}{\partial \sigma_{i'}}\right],$$

[11] More generally, for $k_1, k_2 \in \{0, 1, 2, \ldots\}$ and $j_1, j_2 \in \{0, 1\}$, let

$$\kappa_{k_1,k_2,j_1,j_2} := \mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y \,|\, m, \sigma)}\,\frac{\partial^{k_1+j_1}\ell(Y \,|\, m, \sigma)}{(\partial m)^{k_1}(\partial\sigma)^{j_1}}\,\frac{1}{\ell(Y \,|\, m, \sigma)}\,\frac{\partial^{k_2+j_2}\ell(Y \,|\, m, \sigma)}{(\partial m)^{k_2}(\partial\sigma)^{j_2}}\right]$$
$$= \int_{-\infty}^{\infty}\frac{1}{\ell(y \,|\, m, \sigma)}\,\frac{\partial^{k_1+j_1}\ell(y \,|\, m, \sigma)}{(\partial m)^{k_1}(\partial\sigma)^{j_1}}\,\frac{\partial^{k_2+j_2}\ell(y \,|\, m, \sigma)}{(\partial m)^{k_2}(\partial\sigma)^{j_2}}dy.$$

One then finds

$$\kappa_{k_1,k_2,j_1,j_2} = \mathbb{1}\Big\{k_1 + 2j_1 = k_2 + 2j_2\Big\}\,(k_1 + 2j_1)!\,\sigma^{-[2(k_1+2j_1)-j_1-j_2]}.$$

where $i' \in \{1, \ldots, d\}$ *in the last line. Then,*

$$\kappa(j,k) = \mathbb{1}\Big\{j = k\Big\} \prod_{i=1}^{d} \frac{j_i!}{\sigma_i^{2j_i}}\,, \qquad \rho(j,i) = \begin{cases} \dfrac{2}{\sigma_i^3} & \text{if } j_i = 2, \text{ and all other entries of } j \text{ are zero,} \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 4 is an immediate corollary of Lemma 3. Using the independence of the components of Y we find

$$\frac{\partial^{j^*} \ell(y \,|\, m, \sigma)}{\prod_{i=1}^{d} (\partial m_i)^{j_i}} = \prod_{i=1}^{d} \frac{\partial^{j_i} \ell(y_i \,|\, m_i, \sigma_i)}{(\partial m_i)^{j_i}},$$

and

$$\kappa(j,k) = \prod_{i=1}^{d} \kappa_{j_i, k_i}.$$

Plugging in the result for κ_{jk} in Lemma 3 then gives the result for $\kappa(j,k)$ in Lemma 4. Analogously for $\rho(j,i')$.

D.3 Nonlinear regression with normal errors

Consider the model

$$Y_i = m(X_i; \theta, \eta) + \sigma(X_i; \theta)\, U_i, \qquad U_i \,|\, X \sim iid \mathcal{N}(0,1), \qquad i = 1, \ldots, d,$$

where $m(\cdot;\cdot,\cdot)$ and $\sigma(\cdot;\cdot)$ are known functions, and θ and η are unknown parameters. Ignore θ for the moment and write

$$Y_i = m_i(\eta) + \sigma_i\, U_i.$$

Let $m = (m_1, \ldots, m_d)$ and $\sigma = (\sigma_1, \ldots, \sigma_d)$. The likelihood for $y = (y_1, \ldots, y_d)$ is then given by

$$\ell(y \,|\, \eta) = \ell(y \,|\, m(\eta), \sigma),$$

where $\ell(y \,|\, m, \sigma)$ is given in Lemma 4. Let ∇_η^p be the vector operator that collects all unique derivatives with respect to η up to order p. Let ∇_m^p be the vector operator that collects all unique derivatives with respect to m up to order p. Then, there exists a matrix valued function $M(\eta)$, which only depends on η and on the function $m(\eta)$,

such that

$$\nabla_\eta^p \, \ell(y \,|\, \eta) = M(\eta) \, \nabla_m^p \, \ell(y \,|\, m(\eta), \sigma). \tag{D.2}$$

We want to calculate

$$\mathbb{E}_\eta \left[\frac{\nabla_\eta^p \ell(Y \,|\, \eta)}{\ell(Y \,|\, \eta)} \, \frac{\nabla_\eta^p \ell(Y \,|\, \eta)^\top}{\ell(Y \,|\, \eta)} \right].$$

Lemma 4 gives us explicit expressions for all the components of

$$\mathbb{E}_{m,\sigma} \left[\frac{\nabla_m^p \ell(Y \,|\, m, \sigma)}{\ell(Y \,|\, m, \sigma)} \, \frac{\nabla_m^p \ell(Y \,|\, m, \sigma)^\top}{\ell(Y \,|\, m, \sigma)} \right].$$

Using (D.2) we have

$$\begin{aligned} &\mathbb{E}_\eta \left[\frac{\nabla_\eta^p \ell(Y \,|\, \eta)}{\ell(Y \,|\, \eta)} \, \frac{\nabla_\eta^p \ell(Y \,|\, \eta)^\top}{\ell(Y \,|\, \eta)} \right] \\ &= M(\eta) \, \mathbb{E}_{m(\eta),\sigma} \left[\frac{\nabla_m^p \ell(Y \,|\, m(\eta), \sigma)}{\ell(Y \,|\, m(\eta), \sigma)} \, \frac{\nabla_m^p \ell(Y \,|\, m(\eta), \sigma)^\top}{\ell(Y \,|\, m(\eta), \sigma)} \right] M(\eta)^\top. \end{aligned}$$

Thus, by combining Lemma 4 with the multidimensional Faà di Bruno's formula we get explicit expressions for all the matrices we need.

D.4 Proof of Lemma 3

We already introduced $\phi(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ and $\phi^{(j)}(y) = \frac{d^j \phi(y)}{dy^j}$ above. Let $j, k \in \{0, 1, 2, 3, \ldots\}$. The well-known orthogonality property of Hermite polynomials in (D.1) can be rewritten as

$$\int_{-\infty}^{\infty} \frac{\phi^{(j)}(y) \, \phi^{(k)}(y)}{\phi(y)} \, dy = \mathbb{1}\Big\{ j = k \Big\} \, j! \, . \tag{D.3}$$

Another well-known property of Hermite polynomials is the recurrence relation $h_{j+1}(y) = y h_j(y) - \frac{d}{dy} h_j(y)$. Using this, it is easy to show that for $j > k$ we have

$$\int_{-\infty}^{\infty} \frac{y \, \phi^{(j)}(y) \, \phi^{(k)}(y)}{\phi(y)} \, dy = -\, \mathbb{1}\Big\{ j = k + 1 \Big\} \, j! \, . \tag{D.4}$$

Next, we have

$$\ell(y \,|\, m, \sigma) = \frac{1}{\sigma}\, \phi\left(\frac{y-m}{\sigma}\right), \qquad \frac{\partial^j \ell(y \,|\, m, \sigma)}{(\partial m)^j} = \frac{(-1)^j}{\sigma^{j+1}}\, \phi^{(j)}\left(\frac{y-m}{\sigma}\right).$$

Using this we obtain

$$\begin{aligned}
\kappa_{jk} &:= \mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y \,|\, m, \sigma)} \frac{\partial^j \ell(Y \,|\, m, \sigma)}{(\partial m)^j} \frac{1}{\ell(Y \,|\, m, \sigma)} \frac{\partial^k \ell(Y \,|\, m, \sigma)}{(\partial m)^k}\right] \\
&= \int_{-\infty}^{\infty} \frac{1}{\ell(y \,|\, m, \sigma)} \frac{\partial^j \ell(y \,|\, m, \sigma)}{(\partial m)^j} \frac{\partial^k \ell(y \,|\, m, \sigma)}{(\partial m)^k}\, dy \\
&= \frac{(-1)^{j+k}}{\sigma^{j+k+1}} \int_{-\infty}^{\infty} \frac{1}{\phi\left(\frac{y-m}{\sigma}\right)} \phi^{(j)}\left(\frac{y-m}{\sigma}\right) \phi^{(k)}\left(\frac{y-m}{\sigma}\right) dy \\
&= \frac{(-1)^{j+k}}{\sigma^{j+k}} \int_{-\infty}^{\infty} \frac{\phi^{(j)}(y)\, \phi^{(k)}(y)}{\phi(y)}\, dy \\
&= \mathbb{1}\Big\{j = k\Big\}\, \frac{j!}{\sigma^{2j}}\,,
\end{aligned}$$

where the second to last step employs a change of variables in the integral ($\frac{y-m}{\sigma} \mapsto y$), and the last step uses (D.3). Similarly, for

$$\frac{\partial \ell(y \,|\, m, \sigma)}{\partial \sigma} = -\frac{1}{\sigma^2}\, \phi\left(\frac{y-m}{\sigma}\right) - \left(\frac{y-m}{\sigma^3}\right) \phi^{(1)}\left(\frac{y-m}{\sigma}\right),$$

one finds

$$\begin{aligned}
\rho_j &:= \mathbb{E}_{m,\sigma}\left[\frac{1}{\ell(Y \,|\, m, \sigma)} \frac{\partial^j \ell(Y \,|\, m, \sigma)}{(\partial m)^j} \frac{\partial \log \ell(Y \,|\, m, \sigma)}{\partial \sigma}\right] \\
&= \int_{-\infty}^{\infty} \frac{1}{\ell(y \,|\, m, \sigma)} \frac{\partial^j \ell(y \,|\, m, \sigma)}{(\partial m)^j} \frac{\partial \ell(y \,|\, m, \sigma)}{\partial \sigma}\, dy \\
&= \frac{(-1)^{1+j}}{\sigma^{2+j}} \int_{-\infty}^{\infty} \frac{1}{\phi\left(\frac{y-m}{\sigma}\right)} \phi^{(j)}\left(\frac{y-m}{\sigma}\right) \left[\phi\left(\frac{y-m}{\sigma}\right) + \left(\frac{y-m}{\sigma}\right) \phi^{(1)}\left(\frac{y-m}{\sigma}\right)\right] dy \\
&= \frac{(-1)^{1+j}}{\sigma^{1+j}} \int_{-\infty}^{\infty} \frac{1}{\phi\left(y\right)} \phi^{(j)}\left(y\right) \left[\phi\left(y\right) + y\, \phi^{(1)}\left(y\right)\right] dy \\
&= \mathbb{1}\Big\{j = 2\Big\} \frac{(-1)}{\sigma^3} \int_{-\infty}^{\infty} \frac{y\, \phi^{(1)}\left(y\right)\, \phi^{(2)}\left(y\right)}{\phi\left(y\right)}\, dy \\
&= \mathbb{1}\Big\{j = 2\Big\} \frac{2}{\sigma^3},
\end{aligned}$$

where we again employed the same change of variables in the integration and also used (D.3) and (D.4).

E Monte Carlo simulation

In this section of the appendix we report on the results of a Monte Carlo experiment. We specify a CES model of team production with log-normal errors, where we take the network structure (i.e., the set $\mathcal{K}$ in (5.3)) as given from the empirical data. We fix the true value of the substitution parameter to $\gamma_0 = 1$, the team size parameter to $\beta_0 = 1$, the log-error variance in teams of size 2 to $\sigma_0^2(2) = 1$, and the variance in teams of size 1 to $\sigma_0^2(1) = 1$. This data generating process is designed to approximate what we found on the empirical data.

We report results based on 300 simulations. In each simulated sample, we estimate the parameters using plug-in method-of-moments and the Neyman-orthogonalized method-of-moments estimates of degree $q = 1$ to $q = 6$. As we did in our empirical study, we compute sample-split preliminary estimates of the author fixed-effects based on all their sole-authored publications except for one, selected at random. However, in the simulation exercise we do not cross-fit the estimators, and simply choose a random selection of sole-authored publications for each author in each Monte Carlo run.

In Tables 3 and 4 we show the median, mean, 2.5% quantile, and 97.5% quantile of each estimate across simulations. Starting with the substitution parameter γ, we see that the plug-in estimator is severely biased, with median and mean biases of -50% (expressed in proportion of the true value). For this parameter, all Neyman-orthogonalized estimators are substantially less biased, with a median bias below 3%, and below 1% for $q \geq 4$. However, in some replications the orthogonalized estimates tend to have large values, which is reflected in a somewhat larger mean bias, ranging from 6% at $q = 2$ down to 1.5% at $q = 5$ and $q = 6$, and quantile bands that are not symmetric around the true value.

Turning next to the team size parameter β, we see that both the plug-in and first-order Neyman-orthogonalized estimators are biased, with a median and mean bias of 4%–6%. All orthogonalized estimates of order $q \geq 2$ are virtually unbiased, both for the mean and the median. Moreover, in this case the quantile bands are symmetric around the true parameter value.

Table 3: Monte Carlo simulation

Substitution γ (true value=1)				
	Median	Mean	2.5%	97.5%
Plug-in	0.5054	0.5089	0.4150	0.5966
$q = 1$	0.9723	0.9864	0.7523	1.3259
$q = 2$	1.0310	1.0625	0.7787	1.5571
$q = 3$	1.0100	1.0408	0.7455	1.5795
$q = 4$	0.9964	1.0227	0.7364	1.5556
$q = 5$	0.9898	1.0161	0.7363	1.4904
$q = 6$	0.9897	1.0154	0.7362	1.4904
Team size β (true value=1)				
	Median	Mean	2.5%	97.5%
Plug-in	1.0578	1.0595	1.0277	1.1003
$q = 1$	1.0459	1.0466	0.9971	1.1038
$q = 2$	1.0029	1.0012	0.9299	1.0605
$q = 3$	1.0007	0.9979	0.9249	1.0577
$q = 4$	1.0013	0.9980	0.9226	1.0590
$q = 5$	1.0007	0.9983	0.9203	1.0566
$q = 6$	1.0007	0.9983	0.9214	1.0565

Notes: 300 simulations.

Table 4: Monte Carlo simulation (continued)

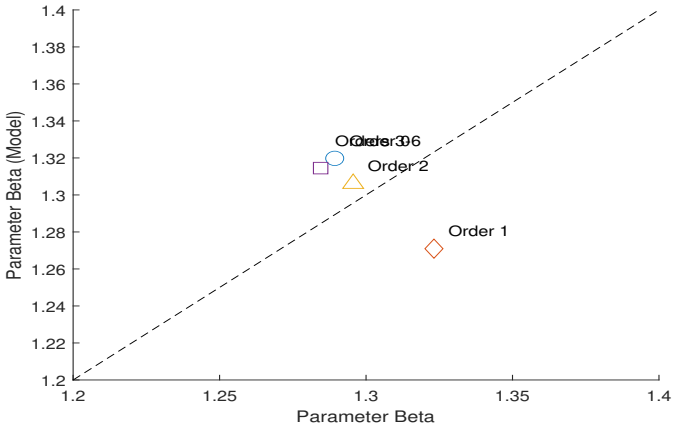
Variance $\sigma^2(2)$ (true value=1)				
	Median	Mean	2.5%	97.5%
Plug-in	1.1623	1.1618	1.1284	1.1944
$q = 1$	1.1824	1.1825	1.1456	1.2228
$q = 2$	0.9964	0.9969	0.9615	1.0334
$q = 3$	1.0001	1.0009	0.9665	1.0366
$q = 4$	0.9995	1.0006	0.9637	1.0373
$q = 5$	0.9993	1.0000	0.9639	1.0365
$q = 6$	0.9982	0.9996	0.9637	1.0374
Variance $\sigma^2(1)$ (true value=1)				
	Median	Mean	2.5%	97.5%
Plug-in	1.3361	1.3368	1.2840	1.3856
$q = 1$	1.2732	1.2723	1.2140	1.3299
$q = 2$	1.0158	1.0178	0.9478	1.0925
$q = 3$	1.0087	1.0119	0.9435	1.0842
$q = 4$	0.9996	1.0033	0.9335	1.0761
$q = 5$	0.9980	1.0002	0.9321	1.0731
$q = 6$	0.9976	0.9992	0.9272	1.0718

Notes: 300 simulations.

Shifting attention to the variance in teams of size 2, $\sigma^2(2)$, we see that both the plug-in and first-order Neyman-orthogonalized estimators are severely biased, with a median and mean bias of 16%–18%. All orthogonalized estimates of order $q \geq 2$ are virtually unbiased, both for the mean and the median, and the quantile bands are centered around the true parameter value.

Lastly, turning to the variance in teams of size 1, $\sigma^2(1)$, the plug-in estimator exhibits a large bias of 33%. First-order orthogonalization only decreases the bias slightly, to 27%. In contrast, the Neyman-orthogonalized estimators continue to show good performance. In particular, when $q \geq 4$ the estimates are virtually unbiased, and the quantile bands are symmetric around the true value.

Figure 1: Comparing two estimates of β_0



Notes: Estimate of β_0 on the x-axis, model-based estimate of β_0 based on the right-hand side of (F.1) on the y-axis. Each point corresponds to an order of orthogonalization.

F Restrictions independent of individual effects

Model (5.3) implies restrictions on parameters $\gamma_0, \beta_0, \sigma_0^2(1), \sigma_0^2(2)$ that do not depend on the author-specific effects η_{i0}.[12] As an example, the model implies the following alternative expression for the team size parameter β_0:

$$\beta_0 = \left(\frac{\mathbb{E}[Y_j^{\gamma_0} \mid s_j = 2]}{\mathbb{E}[Y_j^{\gamma_0} \mid s_j = 1]} \right)^{\frac{1}{\gamma_0}} \exp\left(\frac{1}{2} \gamma_0 [\sigma_0^2(1) - \sigma_0^2(2)] \right), \tag{F.1}$$

where the expectation given $s_j = 1$ is taken over the co-authors' own sole-authored articles, and (F.1) does not involve the fixed-effects η_{i0}. Note that, if $\gamma_0 = 0$ and log-output is additive in worker inputs, then $\log \beta_0$ is simply the difference between average log-outputs in teams of size 2 and 1, respectively. As a check, in Figure 1 we report estimates of the left-hand side of (F.1), against the estimates of β_0 shown in Table 1, for various orders of orthogonalization. We see that the estimates of the two sides of (F.1) tend to agree with each other well irrespective of the orthogonalization order.

Model (5.3) also implies restrictions on γ_0 alone. To see this, let us write (5.3),

[12] The analysis in this section was inspired by discussions with Bo Honoré.

within teams of size 2 only, as

$$Y_j^{\gamma_0} = \frac{1}{2}\beta_0^{\gamma_0}\left(\eta_{k(j,1)0}^{\gamma_0} + \eta_{k(j,2)0}^{\gamma_0}\right)\varepsilon_j^{\gamma_0\sigma_0(2)},$$

which we write in vector form as

$$Y(\gamma_0) = A\widetilde{\eta}_0 + \widetilde{\varepsilon}, \tag{F.2}$$

where $Y(\gamma_0)$ has elements $Y_j^{\gamma_0}$, A is a matrix of zeros and ones,

$$\widetilde{\eta}_{k0} = \frac{1}{2}\beta_0^{\gamma_0}\eta_{k0}^{\gamma_0}\exp\left(\frac{1}{2}\gamma_0^2\sigma_0^2(2)\right),$$

and

$$\widetilde{\varepsilon}_j = \frac{1}{2}\beta_0^{\gamma_0}\left(\eta_{k(j,1)0}^{\gamma_0} + \eta_{k(j,2)0}^{\gamma_0}\right)\left[\varepsilon_j^{\gamma_0\sigma_0(2)} - \exp\left(\frac{1}{2}\gamma_0^2\sigma_0^2(2)\right)\right].$$

Since $\mathbb{E}[\widetilde{\varepsilon}_j \mid A] = 0$, (F.2) implies the conditional moment equalities

$$\mathbb{E}\left[(I - AA^\dagger)Y(\gamma_0) \mid A\right] = 0, \tag{F.3}$$

which only depend on γ_0.

To use (F.3) for estimation, we rely on a set of instruments. For this purpose, we use interacted preliminary estimates $Z_j = \widehat{\eta}_{k(j,1)}\widehat{\eta}_{k(j,2)}$ for $k(j,1), k(j,2)$ the co-authors of j. Since we assume the preliminary estimates are constructed from an independent sample, we have

$$\mathbb{E}\left[Z'(I - AA^\dagger)Y(\gamma_0)\right] = 0. \tag{F.4}$$

Note these restrictions remain valid when ε_j are not Gaussian or not mutually independent, provided they are independent of A. We use GMM estimation based on (F.4), that is,

$$\widehat{\gamma}^{\text{GMM}} = \underset{\gamma}{\text{argmin}}\,|Z'(I - AA^\dagger)Y(\gamma)|. \tag{F.5}$$

We implement this estimator in the same way we have implemented our Neyman-orthogonalized equations. Specifically, we construct preliminary estimates of author effects using all but one sole-authored paper for each author, where we select the held-out sole-authored paper at random.

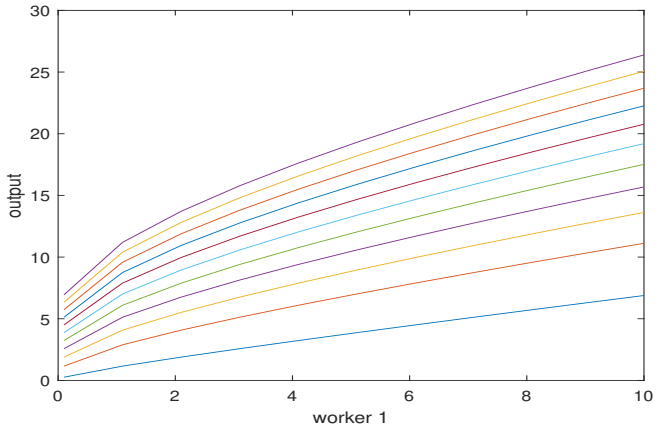
Using the same Monte Carlo simulation design as in Section E tends to give noisy estimates. For example, when the true value is $\gamma_0 = 1$, and $\sigma_0(1) = 1/5$ and

$\sigma_0(2) = 1/5$, we obtain a mean GMM estimate of 1.0218, a median estimate of 0.9725, and a standard deviation of 0.2261 across 300 simulations. Moreover, out of the 300 simulations, in 24 cases we are unable to find another minimum in (F.5) other than $\gamma = 0$ (which is always a minimum by construction). Note these findings correspond to a model with error variances that are *25 times* smaller than the variances we used for our main simulation design in Section E. This suggests this estimation approach, at least for this particular choice of instruments, is considerably less precise than our likelihood-based approach.

Lastly, computing the GMM estimator on the empirical data, cross-fitting 100 times, we obtain $\widehat{\gamma}^{\text{GMM}} = 0.5807$. This is of a comparable magnitude to the estimates of γ reported in Table 1, when using a sufficiently high order of orthogonalization. However, it is worth noting that, out of the 100 random splits of the sole-authored productions, in 5 cases we are unable to find another minimum in (F.5) other than $\gamma = 0$, again reflecting the instability of this method in our setting.

G Estimated production function

Figure 2: Production function estimate



Notes: Worker 1's type η_1 on the x-axis, average output Y_j on the y-axis. Each curve corresponds to a different worker 2's type η_2. Figure based on the point estimates for $q = 6$ reported in Table 1.