Responses to the incidental parameter problem
Pua, A.A.Y.

Citation for published version (APA):
Pua, A. A. Y. (2016). Responses to the incidental parameter problem

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 4

Estimation and inference in dynamic nonlinear fixed effects panel data models by projection

4.1 Introduction

Neyman and Scott (1948) show that the method of maximum likelihood may fail to produce consistent and asymptotically efficient estimators when there are incidental parameters. Lancaster (2000) documents some of the developments after the publication of their paper. Roughly, these developments can be classified into two classes of solutions to the incidental parameter problem: solutions that exploit the structure of the model and solutions that involve orthogonal reparametrization. The latter has been explored more fully in Lancaster (2002) and Woutersen (2003; 2011). Most of the solutions that have been documented are called fixed-$T$ solutions. If one would choose to use an asymptotic scheme where the number of cross-sectional units $n$ grow large, leaving the number of time periods $T$ fixed, then one has to use choose procedures that ensure that the estimating function is both functionally and stochastically independent of the incidental parameters.

Since incidental parameters in panel data models are represented as time-invariant parameters that appear in only a finite number of probability distributions, estimating these parameters induces finite sample bias in the time series dimension. This phenomenon allows us to reconsider the choice of asymptotic scheme. Research by Waterman (1993), Li, Lindsay, and Waterman (2003), and Hahn and Newey (2004)
has paved the way for these large-\(T\) bias corrections. Arellano and Hahn (2007) primarily survey these developments for static panel data models with strictly exogenous regressors. They also document the three related ways of constructing these corrections – correcting the objective function, the moment equation, or the estimator itself. Although one can find consistent estimators of the common parameters, their asymptotic distributions are incorrectly centered. Under this asymptotic scheme, the nonzero center can be estimated when both the number of cross-sectional units and time periods grow at a particular rate (say \(n/T \to c \in (0, \infty)\)). As a result, one can construct an estimator with a correctly centered asymptotic distribution.

In this paper, I adjust the score or some suitably chosen moment function for the common parameter so that a consistent root of the adjusted score has a correctly centered asymptotic distribution. Furthermore, there are cases for which the adjustment produces a fixed-\(T\) consistent estimator. The score or some moment function is the most natural object to adjust because they are the starting points for proofs of consistency and asymptotic normality under regularity conditions. Depending on how one sees the multiple root problem, an issue with score-based adjustments is root selection.\(^1\) In addition, when the common parameter is vector-valued, reconstructing a corrected objective function from the adjusted score or adjusted moment function may no longer be possible.\(^2\) Despite these issues, I discuss some of the advantages of using this score-based adjustment.

First, the computation of the large-\(T\) bias-corrected estimator typically requires the user to select an integer bandwidth whenever a model with some dynamics is being considered. This is true even for the case of a model with lagged dependent variables and strictly exogenous regressors (see for example, Bester and Hansen (2009a) and Hahn and Kuersteiner (2011)) or a static binary choice model with predetermined regressors (see Fernandez-Val (2009)). Arellano and Hahn (2006) modify the objective function which also requires bandwidth selection. The proposed adjustment would not require bandwidth selection just like other score-based corrections (see for example, Woutersen (2003), Carro (2007), and Dhaene and Jochmans (2015b)). One can consider this as an improvement because score-based adjustments exploit the model structure fully in order to create the correction. As a result, finite sample performance may improve, especially in short panels.\(^3\)


\(^2\) One can only recover a quasi-likelihood function from a quasi-score function if the quasi-score is a conservative vector field (see Sections 6.4 and 6.5 of McLeish and Small (1994) for more details). The integration required to go from quasi-score to quasi-likelihood may be path dependent leading to nonuniqueness. The main requirement for a conservative vector field is the symmetry of the derivative matrix of the score. Examples where the latter is not satisfied is in the modelling of covariance matrices in longitudinal data (see Firth and Harris (1991)). It turns out that the symmetry is also required in the context of deriving an information-orthogonal reparameterization. See Section 3.2 of Lancaster (2002).

\(^3\) The score-based adjustment to be discussed later requires the calculation of expectations based on the assumed parametric model. One can avoid the calculation of these expectations by using sample
Second, the approach can accommodate multiple individual-specific fixed effects. Multiple fixed effects may arise when the thresholds in ordered choice models are individual-specific in addition to accounting for individual-specific effects in the linear predictor (see Bester and Hansen (2009a) and Carro and Traferri (2012)). They also arise when a model explicitly allows for a vector of individual-specific effects. For example, Hausman and Pinkovskiy (2013) approximate a dynamic nonlinear model with general predetermined regressors and a scalar individual-specific effect by a Taylor series expansion around an estimator for the scalar individual-specific effect. They show that the transformed model is an affine function of a vector of fixed effects. The elements of this vector are the positive integer powers of the deviation of the scalar individual-specific effect from its estimator. Multiple fixed effects also arise when a model contains time dummies. I do not consider this case but Fernandez-Val and Weidner (2013) have recently proposed and justified the large-$T$ bias corrections in this context.

Third, the approach can accommodate predetermined regressors aside from lagged dependent variables. The approach considered in this paper can accommodate predetermined regressors provided that the feedback process is specified to some degree. The feedback process can either be structural or be some flexible reduced form in the spirit of the Mundlak-Chamberlain device. The specification of the feedback process is partly a matter of interpretation. The Mundlak-Chamberlain device is a correlated random effects approach where the individual-specific fixed effect is usually expressed as a linear projection of the individual-specific fixed effect on the observable characteristics of the cross-sectional unit and a residual (see Mundlak (1978) and Chamberlain (1984)). As proposed by Wooldridge (2000) and applied by Moral-Benito (2013; 2014), the Mundlak-Chamberlain device can be used to flexibly specify the feedback process. In contrast to Wooldridge (2000), we do not specify reduced forms for the individual-specific fixed effect. Corrections that allow for general predetermined regressors without resorting to the device include work by Woutersen (2003), Fernandez-Val (2009), and Fernandez-Val and Weidner (2013).

I give details on the projection approach and its properties in Section 4.2. I also discuss some examples where analytical results are available. In Section 4.3, I present the results of two small-scale Monte Carlo simulations where I compare the projected score to the corrections proposed by Woutersen (2003), Carro (2007), Fernandez-Val (2009), and Hahn and Kuersteiner (2011). Other corrections that were not implemented include the corrections based on (i) modifying the likelihood (see Arellano and Hahn (2006) and Bartolucci et al. (2014)) or integrating the likelihood (see Arellano and Bonhomme (2009) and De Bin, Sartori, and Severini (2015)) and (ii) simulation (see Kim and Sun (2009) and Dhaene and Jochmans (2015b)).
conclude in Section 4.4 and include a technical appendix for some of the calculations and proofs.

4.2 The projection approach

4.2.1 Concept

Suppose we draw a random sample \( \{ y_i = (y_{i1}, \ldots, y_{iT}) : i = 1, \ldots, n \} \) from some known density \( f(y_i; \theta_0, \alpha_{i0}) \), where \( \theta_0 \) is the true value for the common parameter and \( \alpha_{i0} \) is the true value for the incidental parameter. Note that these parameters may be vector-valued but I assume that these are scalars for the purposes of illustration. Denote \( E[\cdot; \theta_0, \alpha_{i0}] \) to be the expectation at the true values of the parameters. Denote \( \partial^k_{\alpha_i} \) to be the \( k \)th order partial derivative with respect to \( \alpha_i \).

To construct consistent estimators for \( \theta_0 \) in the presence of unknown \( \alpha_{i0} \) that have to be estimated, we need a concept that will quantify reduced sensitivity to perturbations of the true value of the incidental parameter, denoted by \( \alpha_i' \), holding \( \theta_0 \) fixed. This means that aside from searching for unbiased estimating functions \( g(\theta, \alpha_i; y_i) \) that have zero expectation at the true value, i.e.

\[
E[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_{i0}] = 0,
\]

we have to further narrow the search to classes of estimating functions that satisfy either of the following conditions:

1. Global ancillarity, where the expectation of the estimating function does not depend on the perturbed value \( \alpha_i' \):

\[
E[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_i'] = 0, \quad \forall \alpha_i' \neq \alpha_{i0}, \tag{4.2.1}
\]

2. \( r \)th-order local \( E \)-ancillarity, where the expectation of the estimating function does not depend on the perturbed value \( \alpha_i' \) within some neighborhood of \( \alpha_{i0} \):

\[
\partial^k_{\alpha_i}E[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_i'] \bigg|_{\alpha_i' = \alpha_{i0}} = 0, \quad \text{for } k = 1, \ldots, r \tag{4.2.2}
\]

Moment functions satisfying (4.2.1) are difficult to construct. Bonhomme (2012) provides a theory that characterizes such moment functions using functional differencing, which is motivated by the theory of orthogonal projections. He also shows that fixed-\( T \) consistent estimation is possible in fully parametric and static and some dynamic panel data settings under some conditions on the distribution of the incidental parameters. Global ancillarity is also equivalent to what Cox and Reid (1987)
call global orthogonality. Tibshirani and Wasserman (1994) call this exact orthogonality in expectation. Woutersen (2011) calls this a zero-score property which holds not just at the true value \( \alpha_{i0} \). Therefore, a sample analog of the score will produce a consistent root regardless of the value plugged in for the incidental parameter.

A more attainable goal is to consider (4.2.2) so that (4.2.1) holds in a smaller region of the parameter space. To further motivate this condition, I expand, up to the second order, the density \( f \) in the left hand side of (4.2.1), i.e.,

\[
\mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha'_i\right] = \int g(\theta_0, \alpha_{i0}; y_i) f(y_i; \theta_0, \alpha'_i) \, dy_i
\]

\[
= \int g(\theta_0, \alpha_{i0}; y_i) f(y_i; \theta_0, \alpha_{i0}) \, dy_i
\]

\[
+ \int g(\theta_0, \alpha_{i0}; y_i) \partial\alpha'_i f(y_i; \theta_0, \alpha'_i) \bigg|_{\alpha'_i = \alpha_{i0}} (\alpha'_i - \alpha_{i0}) \, dy_i
\]

\[
+ \frac{1}{2} \int g(\theta_0, \alpha_{i0}; y_i) \partial^2\alpha'_i f(y_i; \theta_0, \alpha'_i) \bigg|_{\alpha'_i = \alpha_{i0}} (\alpha'_i - \alpha_{i0})^2 \, dy_i
\]

\[
= \mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_{i0}\right] + \partial\alpha'_i \mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha'_i\right] \bigg|_{\alpha'_i = \alpha_{i0}} (\alpha'_i - \alpha_{i0})
\]

\[
+ \frac{1}{2} \partial^2\alpha'_i \mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha'_i\right] \bigg|_{\alpha'_i = \bar{\alpha}_i} (\alpha'_i - \alpha_{i0})^2,
\]

where \( \bar{\alpha}_i \) is in between \( \alpha'_i \) and \( \alpha_{i0} \). Since \( g \) is an unbiased estimating function, the term (a) in the preceding derivation is equal to zero. Under first-order local \( \mathbb{E} \)-ancillarity, the term (b) is also equal to zero. As a result, we have

\[
\mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha'_i\right] = o(\alpha'_i - \alpha_{i0}).
\]

Obviously, the extension to \( r \)-th order local \( \mathbb{E} \)-ancillarity will allow us to conclude that

\[
\mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha'_i\right] = o(\alpha'_i - \alpha_{i0})^r.
\]

Notice that more and more smoothness would be required as one increases \( r \).

First-order local \( \mathbb{E} \)-ancillarity is what Cox and Reid (1987) call information orthogonality or local orthogonality when applied to the likelihood setting. They suggest finding a reparametrization so that \( \theta \) and \( \alpha_i \) are information orthogonal. They call the required transformation an orthogonal reparametrization, which means that,

\[\text{Nonsmooth objective functions, especially those that arise in quantile regressions, are not covered by these ancillarity conditions. It is unclear how smoothing these objective functions will affect the bias-reducing properties of these ancillarity conditions.}\]
up to a certain order, estimating \( \alpha_i \) will have minimal impact on consistently estimating \( \theta \). Lancaster (2002) and Woutersen (2011) derive orthogonal reparametrizations for common panel data models such as the static single index model with strictly exogenous regressors and the linear AR(1) dynamic panel data model. Unfortunately, finding an orthogonal reparametrization requires finding a solution (which may not exist) to a system of partial differential equations.

### 4.2.2 Implications

Instead of finding solutions to the system of partial differential equations and applying the reparametrization, we can determine how \( g \) will satisfy (4.2.2). Notice that \( r \)th-order local \( \mathbb{E} \)-ancillarity is equivalent to searching for \( g \) such that the following set of moment conditions will hold:

\[
\mathbb{E} \left[ g(\theta_0, \alpha_{i0}; y_i) V_i^{(k)}(\theta_0, \alpha_{i0}) \right] = 0, \text{ for } k = 1, \ldots, r,
\]

where

\[
V_i^{(k)}(\theta_0, \alpha_{i0}) = \frac{\partial^k f(y_i; \theta_0, \alpha_{i0})}{f(y_i; \theta_0, \alpha_{i0})}
\]

is the \( k \)th element of the so-called Bhattacharyya basis (see the pioneering works by Bhattacharyya (1946; 1947; 1948)).\(^5\) To show the equivalence, write the left hand side of (4.2.3) as

\[
\mathbb{E} \left[ g(\theta_0, \alpha_{i0}; y_i) V_i^{(k)}(\theta_0, \alpha_{i0}) \right] = \int g(\theta_0, \alpha_{i0}; y_i) \frac{\partial^k f(y_i; \theta_0, \alpha_i')}{f(y_i; \theta_0, \alpha_{i0})} \left| \begin{array}{c} \alpha_i' = \alpha_{i0} \\ \partial \end{array} \right| d y_i
\]

\[
= \frac{\partial^k}{\alpha_i'} \int g(\theta_0, \alpha_{i0}; y_i) f(y_i; \theta_0, \alpha_i') \left| \begin{array}{c} \alpha_i' = \alpha_{i0} \\ \partial \end{array} \right| d y_i
\]

\[
= \frac{\partial^k}{\alpha_i'} \mathbb{E} \left[ g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_i' \right] \left| \begin{array}{c} \alpha_i' = \alpha_{i0} \\ \partial \end{array} \right|
\]

\(^5\)The Bhattacharyya basis is a natural basis to use when studying the effects of fluctuations of the incidental parameters (around the true value) on the density \( f(y_i; \theta_0, \alpha_{i0}) \). Consider a perturbation in the incidental parameter from \( \alpha_{i0} \) to \( \alpha_i' \). A Taylor series expansion of \( f \) about \( \alpha_{i0} \) can be written as the following infinite sum

\[
f(y_i; \theta_0, \alpha_i') = f(y_i; \theta_0, \alpha_{i0}) + \partial_a f(y_i; \theta_0, \alpha_{i0})(\alpha_i' - \alpha_{i0}) + \frac{\partial^2_a f(y_i; \theta_0, \alpha_{i0})}{2} (\alpha_i' - \alpha_{i0})^2 / 2 + \cdots.
\]

The likelihood ratio obtained from comparing the perturbed model to the true model can be written as

\[
\frac{f(y_i; \theta_0, \alpha_i')}{f(y_i; \theta_0, \alpha_{i0})} = 1 + \frac{\partial_a f(y_i; \theta_0, \alpha_{i0})}{f(y_i; \theta_0, \alpha_{i0})} (\alpha_i' - \alpha_{i0}) + \frac{1}{2} \frac{\partial^2_a f(y_i; \theta_0, \alpha_{i0})}{f(y_i; \theta_0, \alpha_{i0})} (\alpha_i' - \alpha_{i0})^2 + \cdots
\]

\[
= 1 + V_i^{(1)}(\theta_0, \alpha_{i0})(\alpha_i' - \alpha_{i0}) + \frac{1}{2} V_i^{(2)}(\theta_0, \alpha_{i0})(\alpha_i' - \alpha_{i0})^2 + \cdots.
\]

Relative to the true model, the perturbed model can be “summarized” in terms of an infinite number of basis elements of the form \( V_i^{(k)}(\theta_0, \alpha_{i0}) \).
where the last expression is equal to zero by (4.2.2). Note that whenever an estimating function \( g \) satisfies \( r \)th-order local \( \mathbb{E} \)-ancillarity, it also satisfies \( k \)th-order local \( \mathbb{E} \)-ancillarity for all \( k = 1, \ldots, r - 1 \).

At this point, I will reduce notation by suppressing the arguments \((\theta_0, \alpha_{i0})\). I now show some of the consequences of (4.2.3) when \( r = 2 \). First, note that

\[
\mathbb{E}\left[ \partial_{\alpha_{i0}} g \right] = \partial_{\alpha_{i0}} \mathbb{E}[g] - \mathbb{E}\left[ g V_i^{(1)} \right] = 0, \tag{4.2.5}
\]

which follows from the requirement that \( g \) be an unbiased estimating function and (4.2.3) when \( r = 1 \). Furthermore, another consequence of (4.2.3) when \( r = 2 \) is

\[
\text{Cov}\left(V_i^{(1)}, \partial_{\alpha_{i0}} g\right) = \mathbb{E}\left[ V_i^{(1)} \partial_{\alpha_{i0}} g \right] - \mathbb{E}\left[ V_i^{(1)} \right] \mathbb{E}[\partial_{\alpha_{i0}} g] = \mathbb{E}\left[ V_i^{(1)} \partial_{\alpha_{i0}} g \right] = 0. \tag{4.2.6}
\]

This zero covariance property follows from calculating the derivative of (4.2.3) with respect to \( \alpha_{i0} \):

\[
\partial_{\alpha_{i0}} \mathbb{E}[g V_i^{(1)}] = \mathbb{E}[g V_i^{(2)}] - \mathbb{E}\left[ V_i^{(1)} \partial_{\alpha_{i0}} g \right]. \tag{4.2.7}
\]

Since \( g \) satisfies first-order local \( \mathbb{E} \)-ancillarity, the expression \( \mathbb{E}\left[ g V_i^{(1)} \right] \) on the left hand side is equal to zero. Since \( g \) satisfies second-order local \( \mathbb{E} \)-ancillarity, the first term in the right hand side of (4.2.7) is equal to zero. As a result, the covariance between \( V_i^{(1)} \) and \( \partial_{\alpha_{i0}} g \) is zero whenever \( g \) satisfies second-order local \( \mathbb{E} \)-ancillarity. Finally,

\[
\mathbb{E}\left[ \partial_{\alpha_{i0}}^2 g \right] = \partial_{\alpha_{i0}} \mathbb{E}\left[ \partial_{\alpha_{i0}} g \right] - \mathbb{E}\left[ V_i^{(1)} \partial_{\alpha_{i0}} g \right] = 0, \tag{4.2.8}
\]

which follows from (4.2.5) and (4.2.6).

It is exactly this zero covariance property (4.2.6), along with the consequences of second-order local \( \mathbb{E} \)-ancillarity (4.2.5) and (4.2.8), that mimics the bias reduction that has already been developed in the literature. Estimator-based corrections in the calculation of the bias of some moment function \( u \) follow from (4.2.5) and (4.2.6), that mimics the bias reduction that has already been developed in the literature.
Notice that if we chose a moment function $u_{it}$ such that

$$E[\partial_{\alpha_i} u_{it}] = 0, \quad E[\psi_{it} \partial_{\alpha_i} u_{it}] = 0, \quad E[\partial_{\alpha_i}^2 u_{it}] = 0,$$

the $O(T^{-1})$ bias disappears. These three equations are exactly the consequences of first-order local $E$-ancillarity, the zero-covariance property in (4.2.6), and second-order local $E$-ancillarity, respectively. It is in this sense that starting from local $E$-ancillarity may be more transparent and intuitive when considering bias corrections.

Let us now consider the case of dynamic nonlinear panel data models. In their motivation for their bias correction procedure, Hahn and Kuersteiner (2011) show that the nonzero center of the asymptotic distribution of the uncorrected MLE is

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{E[\partial_{\alpha_i} v_{it}]} \left[ \frac{1}{T} \left( \sum_{t=1}^T v_{it} \right) \left( \sum_{t=1}^T \partial_{\alpha_i} u_{it} \right) \right] - \frac{E[\partial_{\alpha_i}^2 u_{it}]}{2 \left( E[\partial_{\alpha_i} v_{it}] \right)^2} \frac{1}{T} \left( \sum_{t=1}^T v_{it} \right) \right] \right].$$

Once again notice that if we choose a moment function for the common parameters $u_{it}$ that satisfies second-order local $E$-ancillarity, this nonzero center disappears.

In addition to the preceding discussion, the criterion of second-order local $E$-ancillarity is also constructive because we can interpret (4.2.3) in Hilbert space terms, where the expectation operator is the inner product. We can think of (4.2.3) as finding $g$ that is orthogonal to a linear subspace spanned by $(V_{i1}^{(1)}, \ldots, V_{ir}^{(r)})$. This linear subspace represents local effects of incidental parameter fluctuations. An analogous idea appears in linear regression settings so that we can interpret the desired estimating function $g$ as a residual orthogonal to the explanatory variables $(V_{i1}^{(1)}, \ldots, V_{ir}^{(r)})$. This residual is called the $r$th-order projected score. In principle, one can construct the $r$th-order projected score but a lot of the benefits in terms of bias correction can already be reaped at the second order as seen in the preceding discussions.

### 4.2.3 Computation

Let us consider the situation where one has a complete specification of a likelihood for the data. For every $i = 1, \ldots, n$, let $z_i = (y_{i0}, y_{i1}, \ldots, y_{iT}, x_{i1}, \ldots, x_{iT})$ be the data for the $i$th unit and $z = (z_1, \ldots, z_n)$ be the full data. Let $f(z_{it}; \theta, \alpha_i)$ be the density of the data where $\theta \in \mathbb{R}^p$ and $\alpha_i \in \mathbb{R}^q$. Assume the cross-sectional units are independent of each other. The joint density of the observables is given by

$$f(z; \theta, \alpha) = \prod_{i=1}^n f(z_i; \theta, \alpha_i).$$

Note that the density $f(z_i; \theta, \alpha_i)$ is specified such that predetermined regressors can be accommodated. For example, if we let $x_i^t = (x_{i1}, \ldots, x_{iT})$ and $y_i^t = (y_{i0}, y_{i1}, \ldots, y_{iT})$,
we can write \( f(z_i; \theta, \alpha_i) \) as

\[
\begin{align*}
  f(z_i; \theta, \alpha_i) &= f(y_{iT}|x_i^T, y_{i}^{T-1}; \theta, \alpha_i) \times f(x_{iT}|y_i^{T-1}, x_i^{T-1}) \times \ldots \times f(y_{i2}|x_i^2, y_i^1; \theta, \alpha_i) \\
  &\quad \quad \times f(x_{i2}|y_i^1, x_{i1}) \times f(y_{i1}|x_{i1}, y_{i0}; \theta, \alpha_i) \times f(y_{i0}, x_{i1})
\end{align*}
\]

We usually specify parametric models for \( f(y_{iT}|x_i^T, y_{i}^{T-1}; \theta, \alpha_i) \) and treat these models as structural. Flexible reduced forms can then be used to specify the feedback processes \( f(x_{iT}|y_i^{T-1}, x_i^{T-1}) \). These flexible reduced forms can introduce further individual-specific fixed effects different from \( \alpha_i \). Examples can be found in Moral-Benito (2013; 2014). Note that the distribution of the initial values \( f(y_{i0}, x_{i1}) \) can be specified or be left unspecified. If left unspecified, I condition on initial values.

The \( \theta \)-score and \( \alpha_i \)-score can be written as

\[
\begin{align*}
  U_{i,0}(\theta, \alpha_i; z_i) &= \partial_\theta \log f(z_i; \theta, \alpha_i), \\
  V_i(\theta, \alpha_i; z_i) &= \partial_{\alpha_i} \log f(z_i; \theta, \alpha_i).
\end{align*}
\]

Observe that the \( \alpha_i \)-score only uses the time-series observations for the \( i \)th cross-sectional unit and is a function of \( \alpha_i \) and not of \( \alpha_j \) for \( j \neq i \).

When we set \( k = 1 \) in (4.2.4), \( V_i^{(1)} \) coincides with the \( \alpha_i \)-score so that \( V_i^{(1)} = V_i \). The second-order terms \( V_i^{(2)} \) can be written as

\[
V_i^{(2)} = \partial_{\alpha_i} V_i + V_i V_i^T. \tag{4.2.9}
\]

The preceding recurrence relation, which can be generalized to the \( r \)th order, is a consequence of

\[
\partial_{\alpha_i} V_i = \partial_{\alpha_i} \left( \frac{\partial f}{\partial \alpha_i} \right) = \frac{f \times \partial_{\alpha_i} f \times \partial \alpha_i f}{f^2} = \frac{\partial_{\alpha_i} f}{f} = V_i^{(2)} - V_i V_i^T,
\]

which follows from the quotient rule for derivatives. Note that (4.2.9) is a recurrence relation because one can generate \( V_i^{(r)} \) from \( V_i^{(r-1)} \). Define the second-order extended information matrix as

\[
M_{i,2} = \mathbb{E} \left[ \begin{pmatrix} U_{i,0} \\ V_i \\ \text{vec}[V_i^{(2)}] \end{pmatrix} \begin{pmatrix} U_i^T & V_i^T & \text{vec}[V_i^{(2)}]^T \end{pmatrix} \right] = \begin{pmatrix} M_{11,i} & M_{12,i} \\ M_{21,i} & M_{22,i} \end{pmatrix},
\]

where the submatrices are defined as follows:

\[
\begin{align*}
  M_{11,i} &= \mathbb{E} \left[ U_{i,0} U_{i,0}^T \right], \\
  M_{12,i} &= \mathbb{E} \left[ U_{i,0} V_i^T \right], \\
  M_{21,i} &= \mathbb{E} \left[ U_{i,0} \text{vec}[V_i^{(2)}]^T \right], \\
  M_{22,i} &= \mathbb{E} \left[ \text{vec}[V_i^{(2)}] \text{vec}[V_i^{(2)}]^T \right].
\end{align*}
\]
The second-order projected score and its information matrix for the $i$th unit could be expressed as

$$M_{22,i} = \mathbb{E} \left[ \left( \begin{array}{c} V_i \\ \text{vec} \left[ V_i^{(2)} \right] \end{array} \right) \left( V_i^T \quad \text{vec} \left[ V_i^{(2)} \right]^T \right) \right].$$

The second-order projected score and its information matrix for the $i$th unit could be expressed as

$$U_{i,2} = U_{i,0} - M_{12,i} \left( M_{22,i} \right)^{-1} \left( \begin{array}{c} V_i \\ \text{vec} \left[ V_i^{(2)} \right] \end{array} \right), \quad (4.2.10)$$

$$I_{i,2} = M_{11,i} - M_{12,i} \left( M_{22,i} \right)^{-1} M_{21,i}, \quad (4.2.11)$$

where $\left( M_{22,i} \right)^{-1}$ is the Moore-Penrose inverse of $M_{22,i}$. As discussed in the previous subsection, the second-order projected score is really the residual orthogonal to the linear subspace spanned by $\left( V_i^{(1)}, V_i^{(2)} \right)$. Thus, the second-order projected score $U_{i,2}$ makes the $\theta$-score $U_{i,0}$ less sensitive to the presence of the incidental parameters. The second-order projected score and its associated information matrix for the full data can then be computed by summing up $n$ components of the form (4.2.10) and (4.2.11).

As a result of all the preceding discussions, I present the following lemma and a more formal proof in the appendix.

**Lemma 4.2.1.** The second-order projected score $U_{i,2}$ is an unbiased estimating equation that satisfies second-order local $\mathbb{E}$-ancillarity (4.2.2).

In general, the projected score may depend on both $\theta$ and $\alpha_i$. Thus, we have to substitute an estimator for $\alpha_i$ to form a plug-in projected score. The first-order projected score for the $i$th unit can be written as

$$U_{i,1} = U_{i,0} - \mathbb{E} \left( U_{i,0} V_i^T \right) \left[ \mathbb{E} \left( V_i V_i^T \right) \right]^{-1} V_i.$$

Solving $V_i = 0$ gives an estimator for $\alpha_i$ given $\theta$, denoted by $\widehat{\alpha}_i(\theta)$. The plug-in first-order projected score $\sum_i \widehat{U}_{i,1}$ coincides with the profile score for $\theta$. Dhaene and Jochmans (2015b) show that the panel Poisson model and panel exponential duration model have profile scores that have zero expectation. Therefore, the plug-in first-order projected score mimics the behavior of the profile score when applied to these models.

On the other hand, the second-order projected score is given by

$$U_{i,2} = U_{i,0} - \mathbb{E} \left( U_{i,0} V_i^T \right) \left[ \mathbb{E} \left( V_i V_i^T \right) \right]^{-1} V_i$$

$$- \mathbb{E} \left( U_{i,0} \text{vec} \left[ V_i^{(2)} \right] \right) \left[ \mathbb{E} \left( \text{vec} \left[ V_i^{(2)} \right] \text{vec} \left[ V_i^{(2)} \right]^T \right) \right]^{-1} \text{vec} \left[ V_i^{(2)} \right]. \quad (4.2.12)$$

The next two propositions show that the plug-in second-order projected score matches the properties of existing bias corrections.
Proposition 4.2.2. Assume that the conditions for (4.5.2), the conditions of Lemma 4.2.1, and the central limit theorems for \( V_i, V_i^2, \partial \alpha_i U_i, \) and \( \partial^2 \alpha_i U_i^2 \) hold. Then,

\[
E \left( \hat{U}_{i,2} (\theta_0) - U_{i,2} \right) = E \left( \hat{U}_{i,2} (\theta_0) - U_{i,0} \right) = O(T^{-1}).
\] (4.2.13)

Since the second-order projected score \( U_{i,2} \) satisfies second-order local \( \mathcal{E} \)-ancillarity by Lemma 4.2.1, we have (a) \( E \left[ \partial \alpha_i U_{i,2} \right] = 0 \), (b) zero covariance between \( \partial \alpha_i U_{i,2} \) and \( V_i \), and (c) \( E \left[ \partial^2 \alpha_i U_{i,2} \right] = 0 \). These three implications of second-order local \( \mathcal{E} \)-ancillarity are the most crucial reasons why \( U_{i,2} \) already provides much of the bias reduction that existing methods aim to provide, just as sketched in the preceding subsection.

Assume that the system of equations implied by the plug-in second-order projected score has a solution in some neighborhood of the true value \( \theta_0 \). We denote this solution by \( \mathbf{c} \) and it satisfies

\[
\sum_i \hat{U}_{i,2} (\mathbf{c}) = 0.
\]

This solution has an asymptotic distribution that is exactly the asymptotic distribution of the MLE.

Proposition 4.2.3. Under the asymptotic scheme where \( n, T \to \infty \), \( n/T \to c \in (0, \infty) \), and \( n/T^3 \to 0 \), we have

\[
\sqrt{nT} (\hat{\theta} - \theta_0) \overset{d}{\to} N \left( 0, \lim_{n,T \to \infty} \frac{1}{nT} \sum_{i=1}^{n} E \left[ U_{i,0} U_{i,0}^T \right] \right)^{-1}. \] (4.2.14)

4.2.4 Examples

Consider the following examples to demonstrate the calculations and some of the complications (and virtues) that may arise for the plug-in second-order projected score.

Example 4.2.4. (Linear AR(1) dynamic panel data model) Let \( y_{it} = \alpha_i + \rho y_{i,t-1} + \epsilon_{it} \) where \( \epsilon_{it} \sim \text{iid} \ N \left( 0, \sigma^2 \right) \) for all \( i = 1, \ldots, N \) and \( t = 1, 2 \). Note that I do not restrict \( \rho \) so that \( y_{it} \) will be stationary. I condition on \( y_{i0} \) and assume that it is uncorrelated with future realizations of \( \epsilon_{it} \). The MLE for \( \alpha_i \) given \( \rho \) and \( \sigma^2 \) is \( \hat{\alpha}_i (\rho, \sigma^2) = \hat{y}_i - \rho \hat{y}_{i-1} \), where \( \hat{y}_i = (y_{i1} + y_{i2})/2 \) and \( \hat{y}_{i-1} = (y_{i0} + y_{i1})/2 \). After calculating the second-order projected score for this case,\(^6\) we substitute the MLE for \( \hat{\alpha}_i (\rho, \sigma^2) \) and obtain the following system of equations:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\sigma^2 + (y_{i1} - y_{i0})(y_{i2} - y_{i1} - \rho (y_{i1} - y_{i0}))}{2\sigma^2} = 0,
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{-2\sigma^2 + (y_{i2} - y_{i1} - \rho (y_{i1} - y_{i0}))^2}{4\sigma^4} = 0.
\]

\(^6\)Explicit calculations can be found in the appendix.
Eliminating $\sigma^2$ from the preceding system gives

$$\frac{2}{n} \sum_{i=1}^{n} (y_{i1} - y_{i0})(y_{i2} - y_{i1} - \rho (y_{i1} - y_{i0})) + \frac{1}{n} \sum_{i=1}^{n} (y_{i2} - y_{i1} - \rho ((y_{i1} - y_{i0}))^2 = 0.$$  

Simplifying the equation above gives a quadratic equation in $\rho$ of the form $A_n\rho^2 + B_n\rho + C_n = 0$ where

$$A_n = \frac{1}{n} \sum_{i=1}^{n} (y_{i1} - y_{i0})^2,$$

$$B_n = \frac{-2}{n} \sum_{i=1}^{n} [(y_{i2} - y_{i1})(y_{i1} - y_{i0}) + (y_{i1} - y_{i0})^2],$$

$$C_n = \frac{1}{n} \sum_{i=1}^{n} (y_{i2} - y_{i1})^2 + \frac{2}{n} \sum_{i=1}^{n} (y_{i1} - y_{i0})(y_{i2} - y_{i1}).$$

I now show consistency of one of the roots of the quadratic equation. First, assume that $A_n \overset{p}{\rightarrow} A \neq 0$. Since $\text{Cov}(\epsilon_{i2} - \epsilon_{i1}, y_{i1} - y_{i0}) = -\sigma^2$, we must have $B_n \overset{p}{\rightarrow} -2\rho A + 2\sigma^2 - 2\rho A$ and $C_n \overset{p}{\rightarrow} \rho^2 A - 2\rho \sigma^2 + 2\rho A$. By Slutsky’s lemma, we also have $B_n^2 - 4A_nC_n \overset{p}{\rightarrow} 4(\sigma^2 - A)^2$. This means that the quadratic equation will always have real roots. As a result, we have

$$\tilde{\rho} = \frac{-B_n \pm \sqrt{B_n^2 - 4A_nC_n}}{2A_n} \overset{p}{\rightarrow} \rho - \left(\frac{\sigma^2}{A} - 1\right) \pm \left|\frac{\sigma^2}{A} - 1\right|,$$

where we either have $\tilde{\rho} \overset{p}{\rightarrow} \rho$ or $\tilde{\rho} \overset{p}{\rightarrow} \rho - 2(\sigma^2/A - 1)$. The estimator for $\tilde{\sigma}^2$ is given by

$$\tilde{\sigma}^2_n = \frac{-1}{n} \sum_{i=1}^{n} (y_{i1} - y_{i0})(y_{i2} - y_{i1} - \tilde{\rho} n(y_{i1} - y_{i0})), $$

and will only be consistent if $\tilde{\rho}^2$ is consistent. Notice that the roots were obtained without resorting to an iterative procedure unlike the bias correction proposal by Bun and Carree (2005).

Which of the two roots should be chosen? To illustrate, consider the case where we have stationarity. Assume that $y_{i0}$ is drawn from its stationary distribution where $\mathbb{E}(y_{i0}) = \alpha_i/(1 - \rho)$ and $\text{Var}(y_{i0}) = \sigma^2/(1 - \rho^2)$, where $|\rho| < 1$. In this case, $A_n \overset{p}{\rightarrow} 2\sigma^2/(1 + \rho) \neq 0$. As a result, $\sigma^2/A - 1 < 0$. Thus, the consistent root is the smaller root of the quadratic equation. Now, consider the case where $\rho = 1$. Note that the large-$n$ limit of $A_n$ is such that $\sigma^2/A - 1 < 0$ since $y_{i1} - y_{i0} = \alpha_i + \epsilon_{i1}$ implies that $\mathbb{E}(y_{i1} - y_{i0})^2 = \mathbb{E}(\alpha_i + \epsilon_{i1})^2 = \mathbb{E}(\alpha_i^2) + \sigma^2 > \sigma^2$. As a result, the consistent root is still the smaller root of the quadratic equation.
Dhaene and Jochmans (2015a) extensively document the behavior of the resulting likelihood obtained after integrating the adjusted profile score. They have shown that the profile score has a bias that depends only on the common parameters and not on the incidental parameters. The adjusted profile score is then the difference between the profile score and its bias. They also propose a procedure to choose among the multiple critical points of the adjusted likelihood. Extensions of the model that allow for incidental trends can be found in Moon and Phillips (2004), where they also link the second-order projected score to their proposed moment condition.

Allowing for further lags should be straightforward for the projected score because a scalar $p$th order difference equation can be written as a vector first-order difference equation. Therefore, the quadratic equation derived for the AR(1) case is still going to be a quadratic equation with coefficients that are matrices. Allowing for regressors, whether strictly exogenous or predetermined, will not remove the multiple root problem and will have to be examined on a case-by-case basis.

To explore the effect of including a predetermined regressor, consider an extension of the previous example that automatically allows for two individual-specific fixed effects.

**Example 4.2.5.** (Linear panel VAR(1) model) Consider the following structural model for the dynamics of two variables $(y_{it}, x_{it})$:

\[
\begin{align*}
y_{it} &= \phi_{11}y_{i,t-1} + \phi_{12}x_{i,t-1} + \eta_{xi} + \epsilon_{1it} \\
x_{it} &= \phi_{21}y_{i,t-1} + \phi_{22}x_{i,t-1} + \eta_{yi} + \epsilon_{2it}
\end{align*}
\]

where the idiosyncratic errors have the following distribution:

\[
\begin{pmatrix}
\epsilon_{1it} \\
\epsilon_{2it}
\end{pmatrix}
\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22} \end{pmatrix}\right).
\]

for $i = 1, \ldots, n$ and $t = 1, 2$. Assume that (i) $\Sigma$ is positive definite, i.e., $\det(\Sigma) = \sigma_{11}\sigma_{22} - \sigma_{12}^2 > 0$, (ii) the initial observations $(y_{i0}, x_{i0})$ are available, and (iii) the distribution of the fixed effects and initial observations are left unspecified. The structural parameters are $\theta = (\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}, \sigma_{11}, \sigma_{22}, \sigma_{12})$. The MLEs for $\eta_{xi}$ and $\eta_{yi}$ given the other parameters are

\[
\begin{align*}
\widehat{\eta}_{xi} &= \frac{1}{2} (y_{i2} - y_{i1} (\phi_{11} - 1) - \phi_{11} y_{i0} - \phi_{12} x_{i0} - \phi_{12} x_{i1}) \\
\widehat{\eta}_{yi} &= \frac{1}{2} (x_{i2} - \phi_{21} y_{i0} - \phi_{21} y_{i1} - x_{i1} (\phi_{22} - 1) - \phi_{22} x_{i0})
\end{align*}
\]

The explicit calculations for the projected score can be carried out in Mathematica. The expectation of the plug-in first-order projected score for the $i$th cross-sectional
unit has nonzero bias, i.e.
\[ \mathbb{E}(U_{i,1}) = \left( -\frac{1}{2}, 0, 0, -\frac{\sigma_{22}}{2 \text{det}(\Sigma)}, -\frac{\sigma_{11}}{2 \text{det}(\Sigma)}, -\frac{\sigma_{12}}{\text{det}(\Sigma)} \right). \]
Notice that this nonzero bias does not depend on \( \eta_{xi} \) and \( \eta_{yi} \). As a result, this fits into Case 2 of Dhaene and Jochmans (2015b), where the profile score has expectation free of the incidental parameters. Similarly, calculations in Mathematica show that the expectation of the plug-in second-order projected score for the \( i \)th cross-sectional unit has zero bias. ■

Next, I consider a nonlinear model where the score of some conditional likelihood for the model is an unbiased estimating equation.

**Example 4.2.6.** (Static logit model with strictly exogenous regressors) Suppose \( y_{it} | x_{i1}, x_{i2} \sim \text{Bernoulli}(p_{it}) \) with probability of success \( p_{it} = \mathbb{E}(y_{it} | x_{i1}, x_{i2}) = F(\alpha_i + x_{it}' \beta) \) for \( i = 1, \ldots, n \) and \( t = 1, 2 \). Assume that \( F \) is the logistic cdf. For \( j = 0, 1, 2 \), define \( N_j = \{ i : y_{i1} + y_{i2} = j \} \). Following the computations in the appendix, the second-order projected score using all \( i \) can be computed as
\[
\sum_{i=1}^{n} U_{i,2} = \sum_{i \in N_0} y_{i2} + \sum_{i \in N_2} (y_{i2} - 1) + \sum_{i \in N_1} (x_{i2} - x_{i1})^T \left[ y_{i2} - \frac{1}{1 + e^{(x_{i2} - x_{i1})^T \beta}} \right]
\]
\[
= \sum_{i \in N_1} (x_{i2} - x_{i1})^T \left[ y_{i2} - \frac{1}{1 + e^{(x_{i2} - x_{i1})^T \beta}} \right]. \tag{4.2.15}
\]
Note that the individuals in \( N_0 \) and \( N_2 \) have zero contribution to the plug-in second-order projected score. Although the above expression is monotonically decreasing in \( \beta \), there is no closed-form solution to the above estimating equation. Despite this, the plug-in second-order projected score can be shown that this coincides with score of the conditional likelihood formed from the units for which \( y_{i1} + y_{i2} = 1 \). Since Chamberlain (1980) shows that the conditional MLE is \( \sqrt{n} \)-consistent, the same goes for the root of the plug-in second-order projected score.

Arellano and Bonhomme (2009) derive a bias-reducing prior for this model for general \( T \) that removes the \( O(T^{-1}) \) bias. Their Monte Carlo simulations include an estimator where the adjustment was iterated. The simulations indicate that the iterated adjustment will mimic the properties of the conditional score when \( n \) is fixed and \( T \) increased to around 20. In contrast, Dhaene and Jochmans (2015b), who also consider the case of \( T = 2 \), show that the conditional score can be obtained either by an infinite-order profile score adjustment or by rescaling the profile score by the total number of movers. It is unclear whether rescaling will extend to the case where \( T > 2 \). ■
4.3 Simulations

In this section, I show that the finite sample performance of the plug-in second-order projected score is as good as or sometimes better than some existing competitors. I focus on panels with a very small value of $T$ for the following reasons. Panels obtained from developing countries or panels formed from small-scale experiments usually have single-digit $T$. In practice, applied researchers will also use a subset of the data, especially when there are structural breaks in the time series or when the data are unbalanced. Therefore, it seems appropriate to choose small values of $T$ to gauge finite sample performance.

I implement the projected score method and other alternatives using \textit{Mathematica}.\footnote{All \textit{Mathematica} notebooks and \textit{R} code are available upon request.} \textit{Mathematica} allows us to calculate the symbolic representation of the projected score and to compute the roots using the \texttt{FindRoot} command. Thus, the user only needs to specify the likelihood function and modify the code for the situation he considers without recoding the actual expressions of the corrections.\footnote{Coding the actual expressions would take an inordinate number of lines of code and would only be valid for a specific model.} Furthermore, the calculations become much more compact and organized. I use two starting points, namely, the MLEs for the pooled and fixed effects model, for the root-finding algorithm. I use the software \textit{R} to generate the data for the Monte Carlo experiments and to compute the MLEs (using the routine \texttt{glm}) under the pooled and fixed effects model.\footnote{Whenever the MLE does not exist, I take notice of this and I increase the number of replications so that I could attain the target of 5000 replications.} The draws for the individual-specific fixed effects $\alpha_i$ are fixed across 5000 replications.

The implementation exploits the comparative advantages of both \textit{R} and \textit{Mathematica}. \textit{R} can be used to generate samples from a user-specified data generating process and to perform routine estimation procedures, while \textit{Mathematica} can be used to symbolically calculate the adjusted score and find its roots. The coding style in the \textit{Mathematica} notebook allows any end user to do the following:

1. Specify either an objective function or an estimating function based on some parametric model.

2. Use the built-in commands for differentiation and calculation of expectations to produce symbolic representations of the adjustment found in (4.2.12).

3. Import data and estimation results. The data and estimation results can come from any statistical software capable of exporting its outputs to a text file.

4. Use the programmed functions to generate empirical counterparts of the symbolic representations, to calculate roots and produce output for diagnostics, and to generate routine estimation results such as standard errors.
The coding style almost creates the feeling of a built-in package which may attract more users. But the user only has to change the parametric model in the Mathematica notebook whenever the user contemplates changes in the model.

To construct the plug-in second-order projected score, I compute the projected score as discussed in (4.2.12) and use an estimator for \( \alpha_i \). Rather than recompute \( \tilde{\alpha}_i(\theta) \) at every iteration of the root-finding algorithm, I use a linear approximation of \( \tilde{\alpha}_i(\theta) \) suggested by Bellio and Sartori (2003), i.e.,

\[
\tilde{\alpha}_i(\theta) = \tilde{\alpha}_i + j_{\alpha_i}^{-1}(\tilde{\theta}, \tilde{\alpha}_i) j_{\alpha_i,\theta}(\tilde{\theta}, \tilde{\alpha}_i)(\tilde{\theta} - \theta),
\]

where \( j_{\alpha_i} \) and \( j_{\alpha_i,\theta} \) are the corresponding \( (\alpha_i, \alpha_i) \) and \( (\alpha_i, \theta) \) blocks of the observed information matrix

\[
j(\theta, \alpha_i) = \begin{bmatrix} j_{\theta\theta} & j_{\theta\alpha_i} \\ j_{\alpha_i,\theta} & j_{\alpha_i,\alpha_i} \end{bmatrix},
\]

respectively. Other alternatives may be possible, for instance, using penalized likelihood estimator proposed by Firth (1993) and Kosmidis and Firth (2010) or the EM-based estimator proposed by Chen (2014). The idea behind these estimators is to improve the quality and stability of the plug-in values for \( \alpha_i \). These alternatives may be helpful in models where the plug-in values for \( \alpha_i \) are either extreme or even undefined.

The first data generating process I consider is the static probit model. I use the following design adapted from Fernandez-Val (2009) with some modifications. The original design included a stationary AR(1) model with a linear time trend for the exogenous regressor \( x_{it} \). Omitting this feature leads to the following modified specification:

\[
y_{it} | x_{i1}, \ldots, x_{iT}, \alpha_i \sim Ber(p_{it}), \quad p_{it} = \Phi(\alpha_i + \beta_0 x_{it}),
\]

\[
x_{it} \sim \text{iid } N(0,1), \quad \alpha_i \sim \text{iid } N(0,1), \quad x_{it} \perp \alpha_i,
\]

\[
n = 125, \quad T = 4, \quad \beta_0 = 0.5,
\]

where \( \Phi(\cdot) \) is the standard normal CDF. An important thing to note is that the regressor is already independent of the fixed effects.\(^{10}\) I choose this design because I stripped it down to the simplest elements. I already explored the static logit case in an example found in the previous section.

I also compare the performance of the projected score to the uncorrected MLE, the corrected estimator by Fernandez-Val (2009), and the score corrections by Carro (2007) and Woutersen (2003). Table 4.3.1 contains simulation results for the static probit model based on 5000 replications. The results indicate good finite sample performance of the projected score relative to all the other corrections. The Monte Carlo estimate of the bias is almost reduced by 90% relative to the uncorrected MLE.

\(^{10}\)The exogenous regressor \( x \) is redrawn for every replication for all the experiments in this section.
As a result, taking higher-order projections may not be needed as the gains will be marginal relative to computational cost. Furthermore, the standard deviation of the estimator obtained from the projected score is comparable to the standard deviation of the other estimators. The results clearly indicate that score-based corrections may be preferable in terms of RMSE. Although the number of nonconvergent cases is very small relative to the number of replications, I recommend obtaining a log of the iterations produced by the root-finding algorithm when implementing score-based corrections.

Table 4.3.1: Finite sample performance of estimators of $\beta_0$

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean bias</th>
<th>Median bias</th>
<th>Standard deviation</th>
<th>Median AD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrected MLE</td>
<td>0.210</td>
<td>0.203</td>
<td>0.135</td>
<td>0.089</td>
<td>0.723</td>
</tr>
<tr>
<td>Fernandez-Val (2009)</td>
<td>0.162</td>
<td>0.156</td>
<td>0.123</td>
<td>0.081</td>
<td>0.674</td>
</tr>
<tr>
<td>Woutersen (2003)</td>
<td>0.069</td>
<td>0.064</td>
<td>0.099</td>
<td>0.066</td>
<td>0.577</td>
</tr>
<tr>
<td>(12 cases nonconvergent)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carro (2007)</td>
<td>0.071</td>
<td>0.066</td>
<td>0.100</td>
<td>0.066</td>
<td>0.580</td>
</tr>
<tr>
<td>(13 cases nonconvergent)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projected score</td>
<td>0.030</td>
<td>0.025</td>
<td>0.095</td>
<td>0.063</td>
<td>0.538</td>
</tr>
</tbody>
</table>

Note: True value of $\beta_0$ is equal to 0.5. Results are based on 5000 replications.

The second data generating process is the first-order dynamic logit model. Once more, I adapt the design from Fernandez-Val (2009) with some modifications.

\[
y_{it}|y_{i,t-1}, \ldots, y_{i0}, x_{i0}, x_{i1}, \ldots, x_{iT}, \alpha_i \sim Ber(p_{it}), \quad p_{it} = F\left(\alpha_i + \rho_{0}y_{i,t-1} + \beta_0x_{it}\right),
\]

\[
y_{i0}|x_{i0}, x_{i1}, \ldots, x_{iT}, \alpha_i \sim Ber(p_{i0}), \quad p_{i0} = F\left(\alpha_i + \beta_0x_{i0}\right),
\]

\[
x_{it} \sim iid \text{ L}(0, 1), \quad \alpha_i \sim iid \text{ L}(0, 1), \quad x_{it} \perp \alpha_i,
\]

\[
n = 125, T = 3, \beta_0 = 1, \rho_{0} = 0.5.
\]

In this design, $F(\cdot)$ is the logistic CDF and $L(0, 1)$ is the logistic distribution with mean 0 and scale 1. The original design assumes that $x_{it} \sim N\left(0, \frac{\pi^2}{3}\right)$ and the individual-specific fixed effects were generated as an average of the four oldest values of $x_{it}$. I choose to use $L(0, 1)$ because it is quite similar to $N\left(0, \frac{\pi^2}{3}\right)$ but with heavier tails. I condition on $y_{i0}$ instead of using the information from the distribution $y_{i0}|x_{i0}, x_{i1}, \ldots, x_{iT}$ in the likelihood function. For this model, the alternatives are the fixed -T consistent estimator proposed by Honoré and Kyriazidou (2000), the corrected estimators by Fernandez-Val (2009) and Hahn and Kuersteiner (2011), and the score-based corrections by Carro (2007) and Woutersen (2003).

Recall that Hahn and Kuersteiner (2011) obtain a characterization of the nonzero center of the asymptotic distribution of the MLE as discussed in Example 1.2.3. Estimator-based corrections will have to rely on an estimator of this nonzero center. This nonzero center depends on the cross-spectrum of the $\alpha_i$-score and the deriva-
tive of the $\theta$-score with respect to $\alpha_i$ at the zero frequency and the spectrum of the $\alpha_i$-score at the zero frequency. Since the cross-spectrum and spectrum are infinite sums of cross-covariances and covariances, respectively, a feasible procedure would require some lag truncation. As a result, we would require an integer bandwidth of lower order than $T^{1/2}$ for trimming purposes and for the asymptotic theory to hold. Since $T = 3$, I set the bandwidth at values 0, 1, and 2.

In contrast, Honoré and Kyriazidou (2000) propose an estimator based on the maximizer of a likelihood conditioned on the subset of observations for which $\{x_{i2} = x_{i3}\}$. Since this set is a zero probability event given the DGP, a kernel with a corresponding bandwidth is used to give higher weight to observations where $x_{i2}$ is close to $x_{i3}$ and give lower weight to observations otherwise. I use a standard normal kernel for this purpose. Furthermore, I use the optimal bandwidth derived by Honoré and Kyriazidou (2000) which is a constant multiple of $T^{-1/5}$. I set this constant to values 1, 8, and 64 just as Honoré and Kyriazidou (2000) do in their own simulations.

The Monte Carlo results in Table 4.3.2 indicate that score-based corrections are performing quite well relative to estimator-based corrections for the design I consider. The bias of the root of the projected score is almost eliminated for both coefficients of the linear predictor. In contrast, the other score-based estimators are having problems eliminating the bias in the autoregressive coefficient. There seems to be a point at which a higher bandwidth will not improve finite sample performance of estimator-based corrections. In fact, the estimator-based correction by Hahn and Kuersteiner (2011) almost has the same performance as uncorrected MLE when the bandwidth is equal to 2. Furthermore, the dispersion of the corrected estimators is less than half that of the uncorrected MLE with the exception of the correction by Hahn and Kuersteiner (2011). The dispersion of the root of the projected score is more in line with that of the uncorrected MLE.

I also present two power curves in Table 4.3.1 for the projected score in the dynamic logit model. I do not present the results for the competing procedures because the estimated biases are large relative to the estimated standard deviation. The rejection probability of the test $\rho = 0.5$ is almost 5% while that of the test $\beta = 1$ is about 2%. Unfortunately, power is relatively low but this is expected as the asymptotics require a large value for $T$.

It is clear from the Monte Carlo results that the projected score is a competitive alternative to some of the competing bias-reduction procedures (especially with respect to finite sample bias but not in RMSE terms). The biggest downside is the computational time. For the designs considered, setting up of the projected score, the calculation of the root, and the standard error calculations took about 2 to 5 minutes for every replication on a laptop with 8 GB memory and an i7-processor. Even if we exploit parallel processing, the memory requirement is almost too great for all cores to be used all at once, especially when conducting Monte Carlo simulations. The reason for the high memory requirement is in the nature of the correction – a
Table 4.3.2: Finite sample performance of estimators of $\bar{\beta}$ and $\bar{\rho}$

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean bias $\rho$</th>
<th>Median bias $\beta$</th>
<th>Standard deviation $\rho$</th>
<th>Median AD $\beta$</th>
<th>RMSE $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrected MLE</td>
<td>-2.660</td>
<td>0.867</td>
<td>-2.588</td>
<td>0.792</td>
<td>0.902</td>
</tr>
<tr>
<td>Hahn and Kuersteiner (2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth 0</td>
<td>-0.552</td>
<td>-0.759</td>
<td>-1.066</td>
<td>-0.301</td>
<td>3.049</td>
</tr>
<tr>
<td>Bandwidth 1</td>
<td>-0.341</td>
<td>-0.028</td>
<td>-0.723</td>
<td>0.242</td>
<td>2.870</td>
</tr>
<tr>
<td>Bandwidth 2</td>
<td>-1.957</td>
<td>0.865</td>
<td>-1.870</td>
<td>0.783</td>
<td>0.897</td>
</tr>
<tr>
<td>Fernandez-Val (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth 0</td>
<td>-1.994</td>
<td>0.225</td>
<td>-1.984</td>
<td>0.232</td>
<td>0.555</td>
</tr>
<tr>
<td>Bandwidth 1</td>
<td>-1.807</td>
<td>0.217</td>
<td>-1.795</td>
<td>0.226</td>
<td>0.554</td>
</tr>
<tr>
<td>Bandwidth 2</td>
<td>-1.948</td>
<td>0.211</td>
<td>-1.940</td>
<td>0.220</td>
<td>0.552</td>
</tr>
<tr>
<td>Bandwidth 1</td>
<td>0.268</td>
<td>0.550</td>
<td>1.771</td>
<td>0.885</td>
<td></td>
</tr>
<tr>
<td>Bandwidth 8</td>
<td>-0.049</td>
<td>0.126</td>
<td>0.561</td>
<td>0.265</td>
<td></td>
</tr>
<tr>
<td>Bandwidth 64</td>
<td>-0.059</td>
<td>0.131</td>
<td>0.541</td>
<td>0.250</td>
<td></td>
</tr>
<tr>
<td>Woutersen (2003)</td>
<td>-0.183</td>
<td>-0.047</td>
<td>-0.181</td>
<td>-0.052</td>
<td>0.348</td>
</tr>
<tr>
<td>Carro (2007)</td>
<td>-0.505</td>
<td>-0.047</td>
<td>-0.506</td>
<td>-0.056</td>
<td>0.329</td>
</tr>
<tr>
<td>(1 case nonconvergent)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projected score</td>
<td>0.011</td>
<td>0.048</td>
<td>0.004</td>
<td>0.023</td>
<td>1.170</td>
</tr>
<tr>
<td>(24 cases nonconvergent)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The true values are given by $\rho_0 = 0.5$ and $\beta_0 = 1$. Results are based on 5000 replications.
symbolic representation is created and the data are substituted into this representation. Despite these issues, the implementation is very straightforward and would not require us to program new procedures every time we make changes to the model.

Figure 4.3.1: Inference using the projected score for the dynamic logit model

\[ H_0 : \rho = \rho_0 \]
\[ H_0 : \beta = \beta_0 \]

Note: Significance level set at 5% and represented as horizontal line

4.4 Concluding remarks

This paper develops a bias reduction method for the estimators of common parameters of a linear or nonlinear panel data model with individual-specific fixed effects. The past decades saw a spur of research on bias reduction methods. It is easier to see what these methods have in common by considering what is called the projected score. This projected score is calculated by projecting the score vector for the common parameters onto the orthogonal complement of a space characterized by incidental parameter fluctuations.

I show that projected scores reduce the asymptotic bias of the estimators of common parameters in panel data models. Although the projected score has been introduced two decades ago, its widespread use has been hindered by computational issues. Relative to other bias reduction procedures, computation (in terms of processor time and memory) may be prohibitive but programming is less error-prone and more intuitive. I hope that this will encourage applied researchers to use the projected score. Monte Carlo simulations indicate that the bias-reducing properties of the projected score already take effect even for very small sample sizes usually encountered when panel data models are estimated for subsamples. Finally, the applied researcher need not choose a bandwidth anymore.

Future work on practical aspects include extensions to nonsmooth functions arising, say, in quantile regression. In addition, the projection idea has to be modified when one wants to extend to non-likelihood settings and when one wants to include time effects. I intend to pursue these extensions in the future.
4.5 Appendix

Proof of Lemma 4.2.1

To show that $U_{i,2}$ is an unbiased estimating equation, we have to show that $E[U_{i,0}] = 0$, $E[V_i] = 0$, and $E[vec(V_{i}^{(2)})] = 0$. The first two statements follow from the zero-mean property of the scores. Since $E[vec(V_{i}^{(2)})] = vec(E[V_i])$, we have to show that $E[V_{i}^{(2)}] = 0$. Differentiating $E[V_i] = 0$ with respect to $\alpha_i$ gives the desired result. Thus, we have shown that $U_{i,2}$ is an unbiased estimating equation. To show second-order $E$-ancillarity, we can show that (4.2.12) satisfies the moment conditions in (4.2.3) for $k = 1, 2$. This follows by construction. ■

Proof of Proposition 4.2.2

To simplify the exposition, I return to the case where incidental parameter is scalar.

To show (4.2.13), consider a second-order Taylor series expansion of the plug-in second-order projected score for the $i$th individual about the true value $\alpha_{i0}$, i.e.

$$
\widehat{U}_{i,2}(\theta_0) = U_{i,2} + \partial_{\alpha_i} U_{i,2}(\alpha_i(\theta_0) - \alpha_{i0}) + \frac{1}{2} \partial_{\alpha_i}^2 U_{i,2}(\alpha_i(\theta_0) - \alpha_{i0})^2 + O_p(T^{-1/2}).
$$

(4.5.1)

Under regularity conditions for maximum likelihood estimation, the three terms in (4.5.1) are $O_p(T^{1/2})$, $O_p(T^{1/2})$, and $O_p(1)$. The final term is a zero mean $O_p(T^{-1/2})$ term. Note that the first-order conditions used to obtain a plug-in estimator for $\alpha_i$ can be expanded in the following way:

$$
\widehat{V}_i(\theta_0) = V_i + \partial_{\alpha_i} V_i(\alpha_i(\theta_0) - \alpha_{i0}) + O_p(1).
$$

Since the right hand side is equal to zero, we can write

$$
\alpha_i(\theta_0) - \alpha_{i0} = -\frac{V_i}{\mathbb{E}(\partial_{\alpha_i} V_i)} + O_p(T^{-1}).
$$

(4.5.2)

Furthermore, the square of $\alpha_i(\theta_0) - \alpha_{i0}$ can be written as

$$
(\alpha_i(\theta_0) - \alpha_{i0})^2 = \frac{V_i^2}{[\mathbb{E}(\partial_{\alpha_i} V_i)]^2} - 2 \frac{V_i}{\mathbb{E}(\partial_{\alpha_i} V_i)} O_p(T^{-1}) + O_p(T^{-2})
$$

$$
= \frac{V_i^2}{[\mathbb{E}(\partial_{\alpha_i} V_i)]^2} - 2 \frac{\mathbb{E}(V_i)}{\mathbb{E}(\partial_{\alpha_i} V_i)} O_p(T^{-1}) - 2 \frac{O_p(T^{1/2})}{\mathbb{E}(\partial_{\alpha_i} V_i)} O_p(T^{-1}) + O_p(T^{-2})
$$

93
\[ \text{score has a solution in some neighborhood of the true value} \]

Assume that the system of equations implied by the plug-in second-order projected proof of Proposition 4.2.3

\[ \partial \text{ of the term} \]

The expression in (4.5.5) involves \( \partial \text{ second-order local} \) \( \partial \text{ The expression in (4.5.4) involves the product of} \)

Next, we substitute (4.5.3) into \( \partial, U_{i,2} (\alpha_i (\theta_0) - \alpha_{i0}) \), we have

\[ \partial, U_{i,2} (\alpha_i (\theta_0) - \alpha_{i0}) = - \frac{V_i \partial, U_{i,2}}{\partial, V_i} + \partial, U_{i,2} O_p (T^{-1}) \]

\[ = - \frac{V_i \partial, U_{i,2}}{\partial, V_i} + \partial, U_{i,2} O_p (T^{-1}) + O_p (T^{-1/2}) \]

\[ = - \frac{V_i \partial, U_{i,2}}{\partial, V_i} + O_p (T^{-1/2}) \quad (4.5.4) \]

A central limit theorem for \( \partial, U_{i,2} \) and second-order local \( \partial \text{-ancillarity} \) allow us to produce the previous derivation. The expression in (4.5.4) involves the product of \( \partial, U_{i,2} \) and \( V_i \) and a zero mean \( O_p (T^{-1/2}) \) term. As a result, the expectation of the term \( \partial, U_{i,2} (\alpha_i (\theta_0) - \alpha_{i0}) \) is \( O (T^{-1}) \).

Next, we substitute (4.5.3) into \( \partial^2, U_{i,2} (\alpha_i (\theta_0) - \alpha_{i0})^2 \). As a result, we obtain

\[ \partial^2, U_{i,2} (\alpha_i (\theta_0) - \alpha_{i0})^2 \]

\[ = \frac{\partial^2, U_{i,2} E(V_i^2)}{[E(\partial, V_i)]^2} + \partial^2, U_{i,2} O_p (T^{-3/2}) \]

\[ = \frac{\partial^2, U_{i,2} E(V_i^2)}{[E(\partial, V_i)]^2} + E(\partial^2, U_{i,2}) O_p (T^{-3/2}) + O_p (T^{1/2}) O_p (T^{-3/2}) \]

\[ = \frac{\partial^2, U_{i,2} E(V_i^2)}{[E(\partial, V_i)]^2} + O_p (T^{-1}) \quad (4.5.5) \]

The expression in (4.5.5) involves \( \partial^2, U_{i,2} \), which has zero expectation because of second-order local \( \partial \text{-ancillarity} \), and an \( O_p (T^{-1}) \) term. As a result, the expectation of the term \( \partial^2, U_{i,2} (\alpha_i (\theta_0) - \alpha_{i0}) \) is \( O (T^{-1}) \).

**Proof of Proposition 4.2.3**

Assume that the system of equations implied by the plug-in second-order projected score has a solution in some neighborhood of the true value \( \theta_0 \). We denote this

94
solution by $\hat{\theta}^c$ and it satisfies $\sum_{i=1}^{N} U_{i,2}(\hat{\theta}^c) = 0$. Consider the following first-order Taylor series expansion of the plug-in second-order projected score around $\theta_0$, i.e.

$$\sum_{i=1}^{n} U_{i,2}(\hat{\theta}^c) = \sum_{i=1}^{n} U_{i,2}(\theta_0) + \sum_{i=1}^{n} \frac{d}{d\hat{\theta}} U_{i,2}(\hat{\theta})(\hat{\theta}^c - \theta_0).$$

(4.5.6)

Note that the left hand side of (4.5.6) is equal to zero because $\hat{\theta}^c$ is the root of the plug-in second-order projected score. Rewrite (4.5.6) as

$$\sqrt{n}T (\hat{\theta}^c - \theta_0) = \left( \frac{1}{nT} \sum_{i=1}^{n} \frac{d}{d\hat{\theta}} U_{i,2}(\hat{\theta}) \right)^{-1} \frac{1}{\sqrt{n}T} \sum_{i=1}^{n} U_{i,2}(\theta_0).$$

(4.5.7)

Let $n, T \to \infty$ and $n/T \to c \in (0, \infty)$. Note that

$$\frac{1}{\sqrt{n}T} \sum_{i=1}^{n} [U_{i,2}(\theta_0) - U_{i,2}] = \frac{1}{\sqrt{n}T} \sum_{i=1}^{n} E[U_{i,2}(\theta_0) - U_{i,2}] + O_p(1) = O_p\left(\sqrt{\frac{n}{T^3}}\right) + O_p(1)$$

The first equality comes from replacing the empirical mean with an expectation and leaving behind a zero-mean $O_p(1)$ term. The second equality comes from the order calculation in Proposition 4.2.2. Provided that $n/T^3 \to 0$, $\frac{1}{\sqrt{n}T} \sum_{i=1}^{n} U_{i,2}(\theta_0)$ can be approximated by $\frac{1}{\sqrt{n}T} \sum_{i=1}^{n} U_{i,2}$ and the latter quantity is asymptotically normal. A central limit theorem applies to $\frac{1}{\sqrt{n}T} \sum_{i=1}^{n} U_{i,2}$ (similar to the score in likelihood settings), i.e.

$$\frac{1}{\sqrt{n}T} \sum_{i=1}^{n} U_{i,2} \to N \left(0, \lim_{n,T \to \infty} \frac{1}{nT} \sum_{i=1}^{n} E[U_{i,2}U_{i,2}^T] \right).$$

Next, note that

$$\frac{d}{d\hat{\theta}} U_{i,2}(\hat{\theta}) \bigg|_{\hat{\theta} = \hat{\theta}} = \left[ \partial_{\hat{\theta}} U_{i,2}(\theta) + \left( \partial_{a_i} U_{i,2}(\theta) \right) \left( \partial_{\theta} a_i(\theta) \right) \right] \bigg|_{\hat{\theta} = \hat{\theta}}$$

(4.5.8)

by the chain rule. Replacing $U_{i,2}(\hat{\theta})$ with its Taylor series expansion

$$U_{i,2}(\theta) = U_{i,2}(\theta, a_{i0}) + \partial_{a_i} U_{i,2}(\theta, a_{i0})(a_i(\theta) - a_{i0}) + O_p(1)$$

(4.5.9)

and calculating the derivatives in (4.5.8) yields

$$\partial_{\hat{\theta}} U_{i,2}(\theta) = \partial_{\hat{\theta}} U_{i,2}(\theta, a_{i0}) + \partial_{a_i} U_{i,2}(\theta, a_{i0})(a_i(\theta) - a_{i0})$$

$$+ \partial_{a_i} U_{i,2}(\theta, a_{i0})(\partial_{\theta} a_i(\theta)) + O_p(1),$$

(4.5.10)

$$\partial_{a_i} U_{i,2}(\theta) = \partial_{a_i} U_{i,2}(\theta, a_{i0}) + \partial_{a_i}^2 U_{i,2}(\theta, a_{i0})(a_i(\theta) - a_{i0})$$

$$+ \partial_{a_i}^2 U_{i,2}(\theta, a_{i0})\left( \partial_{a_i}^2 a_i(\theta) \right) + O_p(1).$$

(4.5.11)
Taking probability limits, we have the following components:

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \frac{1}{T} \partial_{\theta} U_{i,2} (\theta, \alpha_{i0}) \bigg|_{\theta = \hat{\theta}} = \lim_{n,T \to \infty} \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E} \left[ \partial_{\theta} U_{i,2} \right], \]

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \frac{1}{T} \partial_{\theta,\alpha_{i0}}^{2} U_{i,2} (\theta, \alpha_{i0}) (\bar{\alpha}_{i}(\theta) - \alpha_{i0}) \bigg|_{\theta = \hat{\theta}} = 0, \]

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \frac{1}{T} \partial_{\alpha_{i0}} U_{i,2} (\theta, \alpha_{i0}) \bigg|_{\theta = \hat{\theta}} = 0, \]

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \frac{1}{T} \partial_{\alpha_{i0}}^{2} U_{i,2} (\theta, \alpha_{i0}) (\bar{\alpha}_{i}(\theta) - \alpha_{i0}) \bigg|_{\theta = \hat{\theta}} = 0, \]

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \lim_{T \to \infty} \frac{1}{T} \partial_{\alpha_{i0}} U_{i,2} (\theta, \alpha_{i0}) (\bar{\alpha}_{i}(\theta)) \bigg|_{\theta = \hat{\theta}} = 0. \]

Note that as \( T \to \infty \), we have both \( \hat{\theta} \to \theta_{0} \) and \( \bar{\alpha}_{i}(\hat{\theta}) \to \alpha_{i0} \). The second and fifth equalities follow \( \bar{\alpha}_{i}(\hat{\theta}) \to \alpha_{i0} \) as \( T \to \infty \). The third, fourth, and sixth equalities would follow from the law of large numbers and second-order ancillarity. The \( O_{p}(1) \) terms in (4.5.9), (4.5.10), and (4.5.11) all converge to zero because \( \bar{\alpha}_{i}(\hat{\theta}) \to \alpha_{i0} \) as \( T \to \infty \). We can then conclude that

\[ \lim_{n,T \to \infty} \frac{1}{nT} \sum_{i=1}^{n} \frac{d}{d\theta} \bar{U}_{i,2}(\theta) \bigg|_{\theta = \hat{\theta}} = \lim_{n,T \to \infty} \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E} \left[ \partial_{\theta} U_{i,2} \right]. \]

Notice that \( U_{i,2} \) behaves like \( U_{i,0} \) asymptotically because the correction in (4.2.12) has expectation zero at the true value. As long as the information identity holds, we have \( \mathbb{E} \left[ U_{i,2}U_{i,2}^{T} \right] = \mathbb{E} \left[ U_{i,0}U_{i,0}^{T} \right] = \mathbb{E} \left[ \partial_{\theta} U_{i,0} \right] \). Otherwise, we have the usual sandwich-type asymptotic covariance matrix.

**Alternative proof of (4.2.14)**

In this appendix, we prove the main results in the spirit of the papers by Hahn and Newey (2004) and Hahn and Kuersteiner (2011). We also note some departures from their proof. Let \( F_{i} \) and \( \bar{F}_{i} \) denote the CDF and its empirical counterpart for the ith individual. Define \( F_{i}(\epsilon) = F_{i} + \epsilon \sqrt{T} (\bar{F}_{i} - F_{i}) \) and \( \Delta_{iT} = \sqrt{T} (\bar{F}_{i} - F_{i}) \), where \( \epsilon \in [0, T^{-1/2}] \). We have \( F(\epsilon) = F + \epsilon \sqrt{T} (\bar{F} - F) \) in vector form.

Let \( \alpha_{i}(\theta, F_{i}(\epsilon)) \) and \( \theta(F(\epsilon)) \) be the solutions to the estimating equations below:

\[ \int V_{i}(\theta, \alpha_{i}(\theta, F_{i}(\epsilon)); z_{i}) dF_{i}(\epsilon) = 0 \quad (4.5.12) \]
\[
\sum_{i=1}^{n} \int U_{i,2}(\theta(F(\epsilon)), \alpha_i(\theta(F(\epsilon)), F_i(\epsilon)); z) \, dF_i(\epsilon) = 0 \tag{4.5.13}
\]

The plug-in used for the \(\alpha_i\)'s in the second-order projected score can be written as \(\hat{\alpha}_i(\theta) = \alpha_i(\theta, F_i(T^{-1/2}))\). The root for the plug-in version of the second-order projected score can be written as \(\hat{\theta} = \theta(F(T^{-1/2}))\). On the other hand, the true values can be written as \(\theta_0 = \theta(F(0)) = \theta(F)\) and \(\alpha_{i0} = \alpha_i(\theta, F_i)\).

Expand the functional \(\theta(\hat{F})\) about the true value \(\theta(F)\) up to the third order, i.e.

\[
\theta(\hat{F}) - \theta(F) = \frac{1}{\sqrt{T}} \theta^\epsilon(0) + \frac{1}{2} \left( \frac{1}{\sqrt{T}} \right)^2 \theta^{\epsilon\epsilon}(0) + \frac{1}{6} \left( \frac{1}{\sqrt{T}} \right)^3 \theta^{\epsilon\epsilon\epsilon}(\epsilon) \tag{4.5.14}
\]

where

\[
\theta^\epsilon(0) = \frac{\partial \theta(F(\epsilon))}{\partial \epsilon}|_{\epsilon=0}, \quad \theta^{\epsilon\epsilon}(0) = \frac{\partial^2 \theta(F(\epsilon))}{\partial \epsilon^2}|_{\epsilon=0}, \quad \theta^{\epsilon\epsilon\epsilon}(\epsilon) = \frac{\partial^3 \theta(F(\epsilon))}{\partial \epsilon^3}|_{\epsilon=\epsilon|_{0,T^{-1/2}}} \tag{4.5.15}
\]

Define the object

\[
h_i(\epsilon) = U_{i,2}(\theta(F(\epsilon)), \alpha_i(\theta(F(\epsilon)), F_i(\epsilon))) \tag{4.5.16}
\]

where the dependence on the data is suppressed. Hahn and Newey (2004) and Hahn and Kuersteiner (2011) use \(U_{i,1}\) instead of \(U_{i,2}\). It follows that (4.5.13) can be rewritten as

\[
\frac{1}{n} \sum_{i=1}^{n} \int h_i(\epsilon) \, dF_i(\epsilon) = 0 \tag{4.5.17}
\]

We show that when \(n, T \to \infty\) such that \(n/T \to c \in (0, \infty)\),

\[
\sqrt{nT}(\theta(\hat{F}) - \theta(F)) \overset{d}{\to} N \left(0, \left( \lim_{n,T \to \infty} \frac{1}{nT} \sum_{i=1}^{n} I_i \right)^{-1} \right) \tag{4.5.18}
\]

in the following manner:

1. Differentiate (4.5.17) with respect to \(\epsilon\) twice. The resulting expressions can be decomposed into two terms: a term that requires integration with respect to \(F_i(\epsilon)\) and a term that characterizes the “tail” or the remainder. We have

\[
\frac{1}{n} \sum_{i=1}^{n} \int \frac{d^2 h_i(\epsilon)}{d\epsilon^2} \, dF_i(\epsilon) + \frac{2}{n} \sum_{i=1}^{n} \int \frac{dh_i(\epsilon)}{d\epsilon} \, d\Delta_iT = 0 \tag{4.5.19}
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \int \frac{d^2 h_i(\epsilon)}{d\epsilon^2} \, dF_i(\epsilon) + \frac{2}{n} \sum_{i=1}^{n} \int \frac{dh_i(\epsilon)}{d\epsilon} \, d\Delta_iT = 0 \tag{4.5.20}
\]
2. Compute the total derivatives in the previous equations noting the dependence of $\theta(F(\epsilon))$ and $\alpha_i(\theta(F(\epsilon)), F_i(\epsilon))$ on $\epsilon$.

\[
\frac{dh_i(\epsilon)}{d\epsilon} = \partial_\theta h_i(\epsilon)\partial_\epsilon \theta + \partial_{\alpha_i} h_i(\epsilon)(\partial_\theta \alpha_i)^T \partial_\epsilon \theta + \partial_{\alpha_i} h_i(\epsilon)\partial_\epsilon \alpha_i
\]

\[
\frac{d^2h_i(\epsilon)}{d\epsilon^2} = \partial_\theta h_i(\epsilon)\partial_\epsilon \theta + \partial_{\alpha_i} h_i(\epsilon)(\partial_\theta \alpha_i)^T \partial_\epsilon \theta + \partial_{\alpha_i} h_i(\epsilon)\partial_\epsilon \alpha_i + \partial_\theta h_i(\epsilon)\partial_\epsilon \alpha_i
\]

3. Next, we have to derive $\theta^\epsilon(0)$ and $\theta^{ee}(0)$. This means that we have to evaluate the expressions in (b) at $\epsilon = 0$. Use the definitions of $\theta_0$, $\alpha_{i0}$ and (4.5.16) to rewrite the resulting expressions. As a consequence, we have

\[
\left. \frac{dh_i(\epsilon)}{d\epsilon} \right|_{\epsilon=0} = \left[ \partial_\theta U_{i,2}(\theta_0, \alpha_{i0}) + \partial_{\alpha_i} U_{i,2}(\theta_0, \alpha_{i0})(\partial_\theta \alpha_i(\theta_0, F_i))^T \right] \theta^\epsilon(0) + \partial_{\alpha_i} U_{i,2}(\theta_0, \alpha_{i0})\partial_\theta \alpha_i(\theta_0, F_i) \theta^\epsilon(0) + \partial_{\alpha_i} U_{i,2}(\theta_0, \alpha_{i0})\partial_\epsilon \alpha_i(\theta_0, F_i)
\]

\[
\left. \frac{d^2h_i(\epsilon)}{d\epsilon^2} \right|_{\epsilon=0} = \theta^\epsilon(0) \left[ \partial_\theta^2 U_{i,2}(\theta_0, \alpha_{i0})\theta^\epsilon(0) + \partial_{\alpha_i}^2 U_{i,2}(\theta_0, \alpha_{i0})\partial_\theta \alpha_i(\theta_0, F_i) \theta^\epsilon(0) \right] + \theta^\epsilon(0) \partial_{\alpha_i}^2 U_{i,2}(\theta_0, \alpha_{i0})\partial_\epsilon \alpha_i(\theta_0, F_i) \theta^\epsilon(0) + \partial_{\alpha_i} \alpha_i(\theta_0, F_i) \theta^\epsilon(0) \left[ \partial_\theta^2 U_{i,2}(\theta_0, \alpha_{i0})\theta^\epsilon(0) \right]
\]

\[
+ \partial_\theta \alpha_i(\theta_0, F_i) \theta^\epsilon(0) \left[ \partial_\theta^2 U_{i,2}(\theta_0, \alpha_{i0})\partial_{\alpha_i} \alpha_i(\theta_0, F_i) \theta^\epsilon(0) \right] + \partial_\theta \alpha_i(\theta_0, F_i) \theta^\epsilon(0) \left[ \partial_{\alpha_i}^2 U_{i,2}(\theta_0, \alpha_{i0})\partial_\theta \alpha_i(\theta_0, F_i) \theta^\epsilon(0) \right]
\]

98
\begin{align*}
+ \partial_\theta \alpha_i(\theta_0, F_i) \theta^0(0) \left[ \partial^2_{\alpha_1} U_{i,2}(\theta_0, \alpha_{i0}) \partial_\epsilon \alpha_i(\theta_0, F_i) \right] \\
+ \partial_{\alpha_1} U_{i,2}(\theta_0, \alpha_{i0})(\theta^0(0))^T \partial^2_{\theta} \alpha_i(\theta_0, F_i) \theta^0(0) \\
+ \partial_{\alpha_1} U_{i,2}(\theta_0, \alpha_{i0}) \partial^2_{\theta,\epsilon} \alpha_i(\theta_0, F_i) \theta^0(0) \\
+ \partial_{\alpha_1} U_{i,2}(\theta_0, \alpha_{i0})(\partial_\theta \alpha_i(\theta_0, F_i) \theta^0(0)) \\
+ \partial_{\epsilon} \alpha_i(\theta_0, F_i) \left[ \partial^2_{\theta,\alpha_1} U_{i,2}(\theta_0, \alpha_{i0}) \partial_\theta \alpha_i(\theta_0, F_i) \theta^0(0) \right] \\
+ \partial_{\epsilon} \alpha_i(\theta_0, F_i) \left[ \partial^2_{\alpha_1} U_{i,2}(\theta_0, \alpha_{i0}) \partial_\epsilon \alpha_i(\theta_0, F_i) \right] \\
+ \partial_{\epsilon} U_{i,2}(\theta_0, \alpha_{i0}) \partial^2_{\theta,\epsilon} \alpha_i(\theta_0, F_i) \theta^0(0) \\
+ \partial_{\alpha_1} U_{i,2}(\theta_0, \alpha_{i0}) \partial^2_{\epsilon} \alpha_i(\theta_0, F_i) \tag{4.5.22}
\end{align*}

4. Substitute the above expressions into (4.5.19) and (4.5.20). The first sum in (4.5.19) and (4.5.20) when evaluated at \( \epsilon = 0 \) becomes the expectation with respect to the true values while the second sum becomes a “tail” term characterizing the difference between the realized distribution \( \hat{F}_i \) and the true one \( F_i \). Since \( \theta^0(0) \) and \( \theta^0(0) \) do not depend on the data, they can be treated as constants with respect to the expectation.

5. We need to derive the expressions for the first and second derivatives of \( \alpha_i(\theta, F_i) \) with respect to \( \theta \) and \( \epsilon \). Differentiate (4.5.12) with respect to \( \theta \) and \( \epsilon \). Solve the resulting system of two equations in two unknowns for \( \partial_\theta \alpha_i(\theta, F_i(\epsilon)) \) and \( \partial_\epsilon \alpha_i(\theta, F_i(\epsilon)) \). Next, get the second derivatives of (4.5.12) with respect to \( \theta \) and \( \epsilon \). Solve the resulting system of three equations in three unknowns for \( \partial^2_\theta \alpha_i(\theta, F_i(\epsilon)) \), \( \partial^2_{\theta,\epsilon} \alpha_i(\theta, F_i(\epsilon)) \), and \( \partial^2_\epsilon \alpha_i(\theta, F_i(\epsilon)) \). In effect, we are applying the Implicit Function Theorem and evaluating at \( \epsilon = 0 \) and \( \theta = \theta_0 \). The resulting first derivatives would be

\[
\partial_\theta \alpha_i(\theta_0, F_i) = -\frac{\mathbb{E}(\partial_\theta V_i^{(1)})}{\mathbb{E}(\partial_{\alpha_1} V_i^{(1)})} = O_p(1) \tag{4.5.23}
\]

\footnote{The systems of equations can be found in the appendix of Hahn and Kuersteiner (2011). Refer to pages 1178 and 1181. Solving the system of equations is not as hard as it sounds because the coefficient matrix is diagonal.}
\[
\partial_e \alpha_i(\theta_0, F_i) = -\frac{T^{1/2}}{T} \left( \frac{V_i^{(1)} - \mathbb{E}(V_i^{(1)})}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \right) = O_p(T^{-1}) \tag{4.5.24}
\]

The resulting second derivatives would be

\[
\partial_{\theta_e}^2 \alpha_i(\theta_0, F_i) = -\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ \mathbb{E}(\partial_{\theta, \alpha_i}^2 V_i^{(1)}) + \partial_{\theta} \alpha_i(\theta_0, F_i) \mathbb{E}(\partial_{\theta, \alpha_i} V_i^{(1)})^T \right] - \frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ \mathbb{E}(\partial_{\theta, \alpha_i}^2 V_i^{(1)}) \mathbb{E}(\partial_{\theta} \alpha_i(\theta_0, F_i)) ^T \right] - \frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ \mathbb{E}(\partial_{\theta, \alpha_i}^2 V_i^{(1)}) \partial_\theta \alpha_i(\theta_0, F_i) \right] = O_p(1) \tag{4.5.25}
\]

\[
\partial_{\theta, \theta_e}^2 \alpha_i(\theta_0, F_i) = -\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ \mathbb{E}(\partial_{\theta, \alpha_i}^2 V_i^{(1)}) \partial_\theta \alpha_i(\theta_0, F_i) + T^{1/2} \frac{1}{T} \left( \partial_\theta V_i^{(1)} - \mathbb{E}(\partial_\theta V_i^{(1)}) \right) \right] - \frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ \mathbb{E}(\partial_{\theta, \alpha_i}^2 V_i^{(1)}) \partial_\theta \alpha_i(\theta_0, F_i) \right] - \frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ T^{1/2} \frac{1}{T} \left( \partial_\theta V_i^{(1)} - \mathbb{E}(\partial_\theta V_i^{(1)}) \right) \partial_\theta \alpha_i(\theta_0, F_i) \right] = O_p(T^{-1}) \tag{4.5.26}
\]

\[
\partial_\theta^2 \alpha_i(\theta_0, F_i) = -\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ \mathbb{E}(\partial_{\theta, \alpha_i}^2 V_i^{(1)}) \partial_e \alpha_i(\theta_0, F_i)^2 \right] - \frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} \left[ 2T^{1/2} \frac{1}{T} \left( \partial_e V_i^{(1)} - \mathbb{E}(\partial_e V_i^{(1)}) \right) \partial_\theta \alpha_i(\theta_0, F_i) \right] = O_p(T^{-2}) \tag{4.5.27}
\]

Central limit theorems are applied to \(\partial_{\alpha_i} V_i^{(1)}\) and \(\partial_\theta V_i^{(1)}\), so that the resulting order of magnitude calculations can be obtained.

6. We are now in a position to simplify \(\theta^e(0)\) and \(\theta^{ee}(0)\).

(a) First, we find an expression for \(\theta^e(0)\). Calculate the expectation of every term in (4.5.21) at the true values. Note that \(\theta^e(0)\) do not depend on the data. Further note that (4.5.23) is already a constant while (4.5.24) depends on the data through \(V_i^{(1)}\).\(^{12}\) Second-order \(E\)-ancillarity implies
that $\mathbb{E}(\partial_{\alpha} U_{i,2}(\theta_0, \alpha_{i_0})) = 0$ and $\mathbb{E}(V^{(1)}_i \partial_{\alpha} U_{i,2}(\theta_0, \alpha_{i_0})) = 0$, as in (4.2.7). As a result, the first sum in (4.5.19) is given by

$$\frac{1}{n} \sum_{i=1}^{n} \int \frac{dh_i(0)}{d\epsilon} \ dF_i = \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\partial_{\theta} U_{i,2}(\theta_0, \alpha_{i_0})) \right) \theta^\epsilon(0) \quad (4.5.28)$$

The remaining term in (4.5.19) is given by

$$\frac{1}{n} \sum_{i=1}^{n} \int h_i(\epsilon) \ d\Delta T = \frac{1}{n} \sum_{i=1}^{n} \int U_{i,2}(\theta_0, \alpha_{i_0}) \ d\Delta T = \frac{\sqrt{T}}{n} \sum_{i=1}^{n} \frac{1}{T} \left( U_{i,2}(\theta_0, \alpha_{i_0}) - \mathbb{E}(U_{i,2}(\theta_0, \alpha_{i_0})) \right) \quad (4.5.29)$$

Define $I_i$ as follows, provided integration and differentiation can be interchanged:

$$I_i = \mathbb{E}[(U_{i,2}(\theta_0, \alpha_{i_0}))(U_{i,2}(\theta_0, \alpha_{i_0}))^T] = \mathbb{E}(\partial_{\theta} U_{i,2}(\theta_0, \alpha_{i_0})) \quad (4.5.30)$$

Thus, we have the following expression for $\theta^\epsilon(0)$, whose asymptotic distribution we seek:

$$\theta^\epsilon(0) = \left( \frac{1}{n} \sum_{i=1}^{n} I_i \right)^{-1} \frac{\sqrt{T}}{n} \sum_{i=1}^{n} \frac{1}{T} \left( U_{i,2}(\theta_0, \alpha_{i_0}) - \mathbb{E}(U_{i,2}(\theta_0, \alpha_{i_0})) \right) \quad (4.5.31)$$

Assume that a central limit theorem holds for $U_{i,2}(\theta_0, \alpha_{i_0})$, i.e.

$$\sqrt{nT} \left( \frac{1}{nT} \sum_{i=1}^{n} U_{i,2}(\theta_0, \alpha_{i_0}) \right) \xrightarrow{d} N \left( 0, \lim_{nT \to \infty} \frac{1}{nT} \sum_{i=1}^{n} I_i \right) \quad (4.5.32)$$

As a consequence, we have

$$\sqrt{nT} \frac{1}{\sqrt{T}} \theta^\epsilon(0) = \left( \frac{1}{nT} \sum_{i=1}^{n} I_i \right)^{-1} \left( \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} U_{i,2}(\theta_0, \alpha_{i_0}) \right) \xrightarrow{d} N \left( 0, \left( \lim_{nT \to \infty} \frac{1}{nT} \sum_{i=1}^{n} I_i \right)^{-1} \right) \quad (4.5.33)$$

Therefore,

$$\theta^\epsilon(0) = O_p(n^{-1/2}) \quad (4.5.34)$$

system of equations mentioned in Step 5 and make the order of magnitude calculations explicit to take into account the latter fact.
(b) Next we find an expression for $\theta^{ee}(0)$. Calculate the expectation of every term in (4.5.22) at the true values while noting the orders of magnitude in (4.5.23), (4.5.24), (4.5.25), (4.5.26), (4.5.27), and (4.5.34). The boxed, double-boxed, oval-boxed and unboxed terms in (4.5.22) are $O_p(n^{-1/2})$, $O_p(T^{-1})$, 0, and $O_p(Tn^{-1})$ respectively. The first sum in (4.5.20) can now be written as

$$\frac{1}{n} \sum_{i=1}^{n} \int \frac{d^2 h_i(e)}{de^2} dF_i(e) = \left( \frac{1}{n} \sum_{i=1}^{n} I_i \right) \theta^{ee}(0) + O_p \left( \frac{1}{T} \right) + O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{1}{\sqrt{Tn}} \right)$$

(4.5.35)

After applying central limit theorems for $\partial_{\theta} U_{i,2}(\theta_0, \alpha_{i0})$ and $\partial_{\alpha_i} U_{i,2}(\theta_0, \alpha_{i0})$ and noting the order of magnitude calculations in (4.5.23), (4.5.24), and (4.5.34), the “tail” term in (4.5.20) can now be written as

$$\frac{2}{n} \sum_{i=1}^{n} \int \frac{dh_i(e)}{de} d\Delta_{i,t} = O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{1}{T} \right)$$

(4.5.36)

As a consequence, we have

$$\sqrt{nT} \left( \frac{1}{\sqrt{T}} \right)^2 \theta^{ee}(0) = \left( \frac{1}{nT} \sum_{i=1}^{n} I_i \right)^{-1} \left[ O_p \left( \frac{1}{\sqrt{nT}} \right) + O_p \left( \frac{1}{\sqrt{T^3}} \right) + O_p \left( \sqrt{\frac{n}{T^5}} \right) \right]$$

$$+ \left( \frac{1}{nT} \sum_{i=1}^{n} I_i \right)^{-1} \left[ O_p \left( \frac{1}{\sqrt{nT^2}} \right) + O_p \left( \frac{1}{T^2} \right) \right]$$

Under the conditions that $n, T \to \infty$ and $n/T \to c \in (0, \infty)$, the distribution of $\theta^{ee}(0)$ becomes degenerate at 0.

(c) The last term in the Taylor series expansion (4.5.14) can be shown to be $o_p(1)$. This step mimics the derivation in Hahn and Kuersteiner (2011).

Projected score for the AR(1) linear dynamic panel data model

The model specification is as follows:

$$Y_{i,t-1} = \{y_{i0}, y_{i1}, \ldots, y_{i,t-1}\},$$

$$y_{it} | Y_{i,t-1} \sim \text{iid} N(\alpha_i + \rho y_{i,t-1}, \sigma_i^2), \quad i = 1, \ldots, n; t = 1, \ldots, T$$

(4.5.37)

Assume $y_{i0}$ is available and we calculate expectations conditional on $y_{i0}$ (so that $E[\cdot]$ is the expectation of some expression conditional on $y_{i0}$). Let $u_{it} = y_{it} - \alpha_i - \rho y_{i,t-1}$. The scores for the common parameters $\rho$ and $\sigma_i^2$ and the incidental parameter $\alpha_i$
are given by:

$$U_{i,0}^\rho = \frac{1}{\sigma^2} \sum_{t=1}^{T} u_{it} y_{i,t-1},$$

$$U_{i,0}^{\sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^{T} u_{it}^2,$$

$$V_{i}^{(1)} = \frac{1}{\sigma^2} \sum_{t=1}^{T} u_{it}.$$

To calculate the second-order projected score, we need the following elements:

$$E \left[ (V_{i}^{(1)})^2 \right] = \frac{1}{\sigma^4} E \left[ \sum_{t=1}^{T} u_{it}^2 + 2 \sum_{t=2}^{T} \sum_{s<t} u_{it} u_{is} \right]$$

$$= \frac{1}{\sigma^4} \left[ \sum_{t=1}^{T} \sigma^2 + 2 \sum_{t=2}^{T} \sum_{s<t} E(u_{it} u_{is}) \right]$$

$$= \frac{T}{\sigma^2}.$$

Note that only the second moments of (4.5.37) are used for the above calculation

$$E \left( V_{i}^{(1)} V_{i}^{(2)} \right) = -\frac{T}{\sigma^4} \sum_{t=1}^{T} E(u_{it}) + \frac{1}{\sigma^6} E \left( \sum_{t=1}^{T} u_{it} \right)^3$$

$$= \frac{1}{\sigma^6} E \left[ \left( \sum_{t=1}^{T} u_{it} \right) \left( \sum_{t=1}^{T} u_{it}^2 + 2 \sum_{t=2}^{T} \sum_{s<t} u_{is} u_{it} \right) \right]$$

$$= \frac{1}{\sigma^6} \left[ \sum_{t=1}^{T} E(u_{it}^2) + \sum_{t=1}^{T} \sum_{s<t} E(u_{it}^2 u_{is}) + 2 \sum_{s=1}^{T} \sum_{s<t} u_{is} u_{it} \right]$$

$$= 0.$$

Thus, $V_{i}^{(1)}$ and $V_{i}^{(2)}$ are orthogonal. Note that we used the third moments of (4.5.37) for the preceding calculation

$$E \left( (V_{i}^{(2)})^2 \right) = \frac{T^2}{\sigma^4} - \frac{2T^2}{\sigma^6} E \left( \sum_{t=1}^{T} u_{it} \right)^2 + \frac{1}{\sigma^8} E \left( \sum_{t=1}^{T} u_{it}^2 + 2 \sum_{s=1}^{T} \sum_{t<s} u_{is} u_{it} \right)^2$$

$$= \frac{T^2}{\sigma^4} - \frac{2T^2}{\sigma^6} + \frac{1}{\sigma^8} \left[ \sum_{t=1}^{T} u_{it}^4 + 2 \sum_{t=2}^{T} \sum_{s<t} u_{it}^2 u_{is}^2 + 4 \sum_{t=2}^{T} \sum_{s<t} u_{is} u_{it} \right]$$

$$= -\frac{T^2}{\sigma^4} + \frac{1}{\sigma^4} [3T + T(T-1) + 2(T-1)(T)]$$

$$= \frac{2T^2}{\sigma^4}.$$
We have used fourth moments of (4.5.37) for the preceding calculation

\[
\begin{align*}
\mathbb{E}(U_{i,0}^4 V_i) &= \frac{1}{\sigma^4} \mathbb{E}\left[ \left( \sum_{t=1}^{T} u_{it} y_{i,t-1} \right) \left( \sum_{t=1}^{T} u_{it} \right) \right] \\
&= \frac{1}{\sigma^4} \left[ \sum_{t=1}^{T} \mathbb{E}(u_{it}^2 y_{i,t-1}) + \sum_{t=2}^{T} \sum_{s < t} \mathbb{E}(u_{it} u_{is} y_{i,s-1}) + \sum_{t=2}^{T} \sum_{s < t} \mathbb{E}(u_{it}^2 u_{is} y_{i,s-1}) \right] \\
&= \frac{1}{\sigma^2} \sum_{t=1}^{T} \mathbb{E}(y_{i,t-1}) \\
&= \frac{1}{\sigma^2} \left[ (1 + \rho + \cdots + \rho^{T-1}) y_{i0} + (T - t) \rho^{t-1} \alpha_i \right].
\end{align*}
\]

The last line follows from recursive substitution. Alternatively, we can impose mean stationarity. Note that

\[
\begin{align*}
\mathbb{E}(U_{i,0}^2 V_i) &= -\frac{T}{2\sigma^4} \sum_{t=1}^{T} \mathbb{E}(u_{it}) + \frac{1}{2\sigma^6} \mathbb{E}\left[ \left( \sum_{t=1}^{T} u_{it}^2 \right) \left( \sum_{t=1}^{T} u_{it} \right) \right] \\
&= \frac{1}{2\sigma^6} \mathbb{E}\left[ \sum_{t=1}^{T} \sum_{t=1}^{T} u_{it}^2 u_{is} \right] \\
&= \frac{1}{2\sigma^6} \left[ \sum_{t=1}^{T} \mathbb{E}(u_{it}^3) + \sum_{t=2}^{T} \sum_{s < t} \mathbb{E}(u_{it}^2 u_{is}) + \sum_{t=2}^{T} \sum_{s < t} \mathbb{E}(u_{is}^2 u_{it}) \right] \\
&= 0,
\end{align*}
\]

\[
\begin{align*}
\mathbb{E}(U_{i,0}^2 V_i^{(2)}) &= -\frac{T^2}{2\sigma^4} \frac{T}{2\sigma^6} \mathbb{E}\left[ \sum_{t=1}^{T} u_{it}^2 \right] + \frac{1}{2\sigma^8} \mathbb{E}\left[ \left( \sum_{t=1}^{T} u_{it}^2 \right) \left( \sum_{t=1}^{T} u_{it}^2 \right) + 2 \sum_{t=2}^{T} \sum_{s < t} u_{it} u_{is} \right] \\
&= -\frac{T^2}{2\sigma^4} \frac{T^2}{2\sigma^6} + \frac{1}{2\sigma^8} T(3\sigma^4 + (T-1)\sigma^4) - \frac{T^2}{2\sigma^4} \\
&= \frac{T}{\sigma^4},
\end{align*}
\]

\[
\begin{align*}
\mathbb{E}(U_{i,0}^2 V_i^{(2)}) &= -\frac{T}{\sigma^4} \mathbb{E}\left[ \sum_{t=1}^{T} u_{it}^3 y_{i,t-1} \right] + \frac{1}{\sigma^6} \mathbb{E}\left[ \left( \sum_{t=1}^{T} u_{it}^3 y_{i,t-1} \right) \left( \sum_{t=1}^{T} u_{it}^2 + 2 \sum_{t=2}^{T} \sum_{s < t} u_{it} u_{is} \right) \right] \\
&= \frac{1}{\sigma^6} \mathbb{E}\left[ \sum_{t=1}^{T} \sum_{t=1}^{T} u_{it}^3 y_{i,t-1} + \sum_{t=1}^{T} \sum_{s < t} u_{is}^2 u_{it} y_{i,s-1} + 2 \sum_{t=2}^{T} \sum_{s < t} u_{is} u_{it}^2 y_{i,s-1} + 2 \sum_{t=2}^{T} \sum_{s < t} u_{is} u_{it}^2 y_{i,s-1} \right] \\
&= \frac{2}{\sigma^6} \mathbb{E}\left[ \sum_{t=2}^{T} \sum_{s = t-1}^{T} u_{is} u_{it}^2 y_{i,s} \right].
\end{align*}
\]
Thus, the second-order projected score for an arbitrary value of $T$ can be computed in a straightforward manner using all the components cited above.

Projected score for the static binary choice model with an exogenous regressor

Suppose $y_t|x_1, x_2 \sim Ber(p_t)$ with

$$p_t = \mathbb{E}(y_t|x_1, x_2) = F(\alpha + x_i^T \beta) \equiv F_t$$

for $i = 1, \ldots, n$ and $t = 1, 2$. The uncentered moments of this conditional distribution are all equal to $F_t$. For this discussion, I suppress the dependence of the expression on $i$. Calculations in a separate Mathematica file give the following analytical results specific to the static logit model with one exogenous regressor for $T = 2$. Let

$$D_1 = (e^{\alpha + \beta x_1} + 1)(e^{\alpha + \beta x_2} + 1),$$
$$D_2 = 4e^{\alpha + \beta x_1 + \beta x_2} + e^{2 \alpha + 2 \beta x_1 + \beta x_2} + e^{2 \alpha + \beta x_1 + 2 \beta x_2} + e^{\beta x_1} + e^{\beta x_2}.$$

The scores for $\beta$ and $\alpha$ are given by

$$U_0 = \frac{x_1(e^{\alpha + \beta x_2} + 1)(y_1(e^{\alpha + \beta x_1} + 1) - e^{\alpha + \beta x_1}) + x_2(e^{\alpha + \beta x_1} + 1)(y_2(e^{\alpha + \beta x_2} + 1) - e^{\alpha + \beta x_2})}{D_1},$$
$$V^{(1)} = \frac{-e^{\alpha}(2e^{\alpha + \beta x_1 + \beta x_2} + e^{\beta x_1} + e^{\beta x_2}) + (y_1 + y_2)D_1}{D_1}.$$

The components of the second-order projected score are calculated below. First,

$$\mathbb{E}(V^{(1)})^2 = \frac{e^{\alpha}D_2}{D_1^2}$$

$$\mathbb{E}(V^{(1)}V^{(2)}) = \frac{e^{\alpha}(e^{3 \alpha + \beta x_1 + \beta x_2} - e^{4 \alpha + 3 \beta x_1 + 2 \beta x_2} + e^{3 \alpha + \beta x_1 + 3 \beta x_2} - e^{4 \alpha + 2 \beta x_1 + 3 \beta x_2})}{D_1^3}$$
$$+ \frac{e^{\alpha}(e^{-2 \alpha + \beta x_1} - e^{\alpha + 2 \beta x_2} + e^{\beta x_1} + e^{\beta x_2})}{D_1^3}$$
$$+ \frac{e^{\alpha}(6e^{\alpha + \beta x_1 + \beta x_2} - 6e^{3 \alpha + 2 \beta x_1 + 2 \beta x_2})}{D_1^3}.$$
\[ E(U_0V^{(1)}) = \frac{e^\alpha(x_2e^{\beta x_2}(e^{\alpha+\beta x_2} + 1)^2 + x_1e^{\beta x_1}(e^{\alpha+\beta x_2} + 1)^2)}{D_2^2}. \]

Next, \( V^{(2)} \) is given by

\[
\begin{align*}
V^{(2)}_{y_1=0, y_2=0} &= \frac{e^\alpha \left( e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2} \right)}{D_1^2} \\
&+ \frac{e^\alpha \left( 3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2} \right)}{D_1^2} \\
V^{(2)}_{y_1=1, y_2=0} &= \frac{e^\alpha \left( e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2} \right)}{D_1^2} \\
&+ \frac{e^\alpha \left( 3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2} \right)}{D_1^2} \\
V^{(2)}_{y_1=0, y_2=1} &= \frac{e^\alpha \left( e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2} \right)}{D_1^2} \\
&+ \frac{e^\alpha \left( 3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2} \right)}{D_1^2} \\
V^{(2)}_{y_1=1, y_2=1} &= \frac{e^\alpha \left( e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2} \right)}{D_1^2} \\
&+ \frac{e^\alpha \left( 3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2} \right)}{D_1^2}
\end{align*}
\]

To orthogonalize \( V^{(2)} \), we compute

\[ V^{(2),*} = V^{(2)} - \frac{E(V^{(1)}V^{(2)})}{E((V^{(1)})^2)} V^{(1)}. \]

Depending on the binary patterns of the sequence \((y_1, y_2)\), we have

\[
\begin{align*}
V^{(2),*}_{y_1=0, y_2=0} &= \frac{2(e^{\beta x_1} + e^{\beta x_2}) e^{2\alpha+\beta x_1+\beta x_2}}{D_2}, \\
V^{(2),*}_{y_1=1, y_2=1} &= \frac{2(e^{\beta x_1} + e^{\beta x_2})}{D_2}, \\
V^{(2),*}_{y_1=0, y_2=1} &= -\frac{4e^{\alpha+\beta x_1+\beta x_2}}{D_2}, \\
V^{(2),*}_{y_1=1, y_2=0} &= -\frac{4e^{\alpha+\beta x_1+\beta x_2}}{D_2}.
\end{align*}
\]
Hence, we have

\[ E \left( U_0 V^{(2),*} \right) = \frac{2(x_1 - x_2)(e^{\beta x_2} - e^{\beta x_1})e^{2\alpha + \beta x_1 + \beta x_2}}{D_2}, \]

\[ E(V^{(2),*})^2 = \frac{4(e^{\beta x_1} + e^{\beta x_2})e^{2\alpha + \beta x_1 + \beta x_2}}{D_1 D_2}. \]

The second-order projected score can now be written as

\[ U_2 = \frac{(x_1 - x_2)(y_1^2(e^{\beta x_1} - e^{\beta x_2}) + y_1(-e^{\beta x_1} + 3e^{\beta x_2} + 2y_2(e^{\beta x_1} - e^{\beta x_2})))}{2(e^{\beta x_1} + e^{\beta x_2})} \]
\[ + \frac{(x_1 - x_2)y_2(-3e^{\beta x_1} + e^{\beta x_2} + y_2(e^{\beta x_1} - e^{\beta x_2}))}{2(e^{\beta x_1} + e^{\beta x_2})} \]

Note that we have \( U_2 = 0 \) over cross-sectional units for which \( y_1 + y_2 = 0 \) or \( y_1 + y_2 = 2 \). For cross-sectional units for which \( y_1 + y_2 = 1 \), i.e., substituting in \( y_1 = 1 - y_2 \) in the expression for \( U_2 \), gives the expression one sees in (4.2.15).