Inference in Approximately Sparse Correlated Random Effects Probit Models

Jeffrey M. Wooldridge∗ Ying Zhu†

Updated on July 3, 2017
(Forthcoming in The Journal of Business and Economic Statistics)

Abstract

We propose a simple procedure based on an existing “debiased” $l_1$-regularized method for inference of the average partial effects (APEs) in approximately sparse probit and fractional probit models with panel data, where the number of time periods is fixed and small relative to the number of cross-sectional observations. Our method is computationally simple and does not suffer from the incidental parameters problems that come from attempting to estimate as a parameter the unobserved heterogeneity for each cross-sectional unit. Further, it is robust to arbitrary serial dependence in underlying idiosyncratic errors. Our theoretical results illustrate that inference concerning APEs is more challenging than inference about fixed and low dimensional parameters, as the former concerns deriving the asymptotic normality for sample averages of linear functions of a potentially large set of components in our estimator when a series approximation for the conditional mean of the unobserved heterogeneity is considered. Insights on the applicability and implications of other existing Lasso based inference procedures for our problem are provided. We apply the debiasing method to estimate the effects of spending on test pass rates. Our results show that spending has a positive and statistically significant average partial effect; moreover, the effect is comparable to found using standard parametric methods.

JEL Classification: C14, C23, C25, C55
Keywords: Nonlinear panel data models; Correlated random effects probit; Partial effects; High-dimensional statistics and inference; $l_1$-Regularized Quasi-Maximum Likelihood Estimation

1 Introduction

Probably the most commonly used nonlinear panel data models are those for binary responses $Y_{it} \in \{0, 1\}$, with the most common specification being

$$Y_{it} = 1 \{W_{it} \beta_0 + \alpha_i + \upsilon_{it} > 0\}, \ t = 1, \ldots, T, \ i = 1, \ldots, n,$$

(1)

where $W_{it}$ is a vector of observable covariates, $\alpha_i$ is an unobserved random variable – the unobserved heterogeneity – and $\beta_0$ is a vector of unknown parameters. The $\{\upsilon_{it} : t = 1, \ldots, T\}$ are the idiosyncratic errors for population unit $i$. More recently, many empirical researchers are interested

∗wooldri1@msu.edu; Michigan State University, Department of Economics
†yzhu@msu.edu; Michigan State University, Department of Economics
in estimation models for a fractional response variable $Y_{it} \in [0, 1]$, where $Y_{it}$ may be continuous or take on certain values with positive probability. In this case, we are interested in a conditional mean:

$$E(Y_{it}|W_{it}, \alpha_i) = \Phi(W_{it}\beta_0 + \alpha_i), \quad t = 1, ..., T, \quad i = 1, ..., n.$$  \hspace{1cm} (2)

Under the standard assumptions for the unobserved effects probit model, the conditional mean in (2) also holds in the binary response model (1), and so we can treat both binary and fractional responses in the same framework. We are interested in the case where the number of time periods, $T$, is small relative to the number of cross-sectional observations, $n$. Therefore, our asymptotic analysis will be for fixed $T$ with $n$ increasing to infinity.

Our approach for modeling the unobserved heterogeneity $\alpha_i$ in (2) is based on the concept of correlated random effects (CRE) – as exposited, for example, in Wooldridge (2010, Chapter 13). This approach views the unobserved effect, $\alpha_i$, as a random variable that may be correlated with the history $W_i := \{W_{it} : t = 1, ..., T\}$. Typically, $\alpha_i$ is modeled as $\alpha_i = V_i\gamma_0 + \eta_i$, where $V_i$ is a set of functions of $W_i$ and $\eta_i$ is the heterogeneity left unexplained. In this paper, we consider the case $\eta_i | W_i \sim \mathcal{N}(0, \sigma^2_\eta)$, which makes the conditional mean of the probit form after averaging out $\eta_i$. Compared with the traditional random effects framework, which assumes independence between $\alpha_i$ and $W_i$, the CRE approach is clearly more general.

The CRE approach does not suffer from the incidental parameters problems that come from attempting to estimate as a parameter the unobserved heterogeneity for each cross-sectional unit. Plus, it can be applied to models with substantial heterogeneity, whereas so called “fixed effects” approaches are limited by the number of time periods. Most importantly, the “fixed effects” approach, with large $n$ and small $T$, does not allow us to estimate functions (such as average partial effects) that involve the conditional mean of the heterogeneity. Compared to procedures that require obtaining the joint distribution of the underlying data for each cross-sectional observation over time, the CRE approach in conjunction with a pooled estimation is much simpler computationally. Also, as discussed in Wooldridge (2010), the CRE approach identifies average partial effects (APEs) without any restrictions on the serial dependence in the data. The data can be strongly dependent over time provided the cross-sectional dimension is reasonably large.

The typical approach to CRE models is to specify a simple relationship between $\alpha_i$ and time-constant functions of $\{W_{it}\}$. A leading case is to use a linear function of the time averages, $W_i = T^{-1}\sum_{t=1}^T W_{it}$, which was proposed by Mundlak (1978) in the linear model. Chamberlain (1982) suggested a linear model for the entire history, $W_i$. But these are just simple possibilities, and there are countless others. For example, we could compute variances and covariances of the elements in $W_{it}$. We could break the time interval into, say, $G$ groups, and use statistics computed within the $G$ groups. For flexibility in the functional form, we might want to include polynomials in any set of original functions.

One possible approach to allow for a flexible conditional mean functional form is to use series estimation once a given set of functions of the covariates has been specified. The drawback to this approach is that one must specify which functions of the covariates are to be added as the sample size grows, and no economic theory can provide guidance. In this paper we show how to apply high-dimensional selection methods in a CRE probit model, where we allow substantial flexibility in how the heterogeneity relates to the history of the covariates. However, it is worth emphasizing that our inference procedure and theory does not require perfect selection of our selector.

We must emphasize that this paper only aims to relax the parametric restriction on the conditional mean of the unobserved heterogeneity and does not attempt to relax the probit functional form, as the former is likely the most serious restriction. Simulation results in Li and Zheng (2008) suggest that, for obtaining partial effects, the estimates are not overly sensitive to the normality
normality is derived by conditioning on the covariates

\[ \hat{\theta} \]

ated at the growth condition on the number of non-zero components in Javanmard and Montanari, randomness in both \( \theta \) and \( \Phi \) (based on constructing a “debiased” version of the parameters (e.g., Belloni, Chernozhukov, and Hansen, 2014). Instead, we adopt a different approach partialling out the effects of high-dimensional controls to obtain inference results on low-dimensional approximations for the conditional mean of the unobserved heterogeneity is considered.

This fact renders some of the existing procedures inapplicable, particularly those relying on partialling out the effects of high-dimensional controls to obtain inference results on low-dimensional parameters (e.g., Belloni, Chernozhukov, and Hansen, 2014). Instead, we adopt a different approach based on constructing a “debiased” version of the \( \ell_1 \)-regularized pooled quasi-maximum likelihood estimator. Our procedure is motivated by Javanmard and Montanari (2014), who focus on the linear regression model \( Y_i = X_i\theta^* + u_i \) and inference of a single parameter with cross-sectional data. Given an initial Lasso estimate \( \hat{\theta} \) of \( \theta^* \), Javanmard and Montanari (2014) adds a correction term to \( \hat{\theta} \) to remove the bias introduced by the regularization. A key step in their procedure lies in searching for an approximation \( M \) of the inverse of the population Hessian \( \left[ \mathbb{E} \left( X_i^T X_i \right) \right]^{-1} \) and establishing a finite-sample bound on \( \left| \frac{M}{n} \sum_{i=1}^n X_i^T X_i - I \right|_\infty \) where \( | \cdot |_\infty \) is the elementwise \( l_\infty \)-norm and \( I \) is the identity matrix.

This approximation step becomes more involved in the analysis of a probit-like model \( \mathbb{E} (Y_i | X_i) = \Phi (X_i\theta^*) \) as the inverse of the population Hessian depends on the \( p \)-dimensional coefficient vector \( \theta^* \) (where \( p \) is allowed to exceed \( n \)) and consequently its approximation, \( M \), will inevitably rely on the initial Lasso estimate \( \hat{\theta} \) of \( \theta^* \). Letting \( \hat{H}(\hat{\theta}) \) denote the sample Hessian of our interest evaluated at \( \hat{\theta} \), we use a discretization argument and a metric entropy result to derive a finite-sample bound on \( \left| M \hat{H}(\hat{\theta}) - I \right|_\infty \). In contrast to Javanmard and Montanari (2014) where the asymptotic normality is derived by conditioning on the covariates \( X \), our results incorporate the additional randomness in both \( X \) and \( \hat{\theta} \) as our \( M \) also depends on \( \hat{\theta} \) (clearly in the case of linear models as in Javanmard and Montanari, \( M \) would only depend on \( X \)). It turns out doing so requires imposing a growth condition on the number of non-zero components in \( \hat{\theta} \), \( |J(\hat{\theta})| \).

It is possible to trade off the condition on \( |J(\hat{\theta})| \) with a sparsity assumption on the inverse of the population Hessian by applying a different debiasing method proposed by van de Geer, Bühlmann, Ritov, and Dezeure, (2014) as well as Zhang and Zhang (2014). We choose not to adopt their assumption. However, the findings of Arellano and Bonhomme (2009) and Rabe-Hesketh and Skrondal (2013) suggest that misspecification of the conditional mean can cause substantial bias, even in the partial effects. Relaxing the shape of the distribution is clearly worthwhile and has been done in the work of Altonji and Matzkin (2005). Nevertheless, Altonji and Matzkin assume, at a minimum, that the distribution of the heterogeneity given the history of covariates is exchangeable. Time averages and variances over time satisfy exchangeability, but many functions – such as individual-specific trends – do not. Moreover, the Altonji-Matzkin estimation approach is complicated and not especially appealing to empirical researchers. In comparison, our approach is computationally attractive and does not require an exchangeability assumption.

Our main contribution is to propose valid inference for the average partial effect (APE, or average marginal effect) with respect to any policy variable. If the policy variable of interest indicates the treatment status, the corresponding APE is known as the average treatment effect in literature. When a nonlinear probability model is used to analyze the importance of policy interventions, the APEs are more sensible measures for the magnitudes of effects than the individual parameters themselves (which do not convey much information other than the signs they have). Unlike in a linear model where the APEs coincide with the coefficients, inference of the APEs in a nonlinear model like probit requires establishing distributional results that involve the estimator for the unknown conditional mean function of the unobserved heterogeneity. This problem can be reduced to deriving the asymptotic normality for sample averages of linear functions of a potentially large set of components in our estimator for primary parameters (associated with policy-related variables) as well as nuisance parameters (associated with approximating terms) when a series approximation for the conditional mean of the unobserved heterogeneity is considered.

This fact renders some of the existing procedures inapplicable, particularly those relying on partialling out the effects of high-dimensional controls to obtain inference results on low-dimensional parameters (e.g., Belloni, Chernozhukov, and Hansen, 2014). Instead, we adopt a different approach based on constructing a “debiased” version of the \( \ell_1 \)-regularized pooled quasi-maximum likelihood estimator. Our procedure is motivated by Javanmard and Montanari (2014), who focus on the linear regression model \( Y_i = X_i\theta^* + u_i \) and inference of a single parameter with cross-sectional data. Given an initial Lasso estimate \( \hat{\theta} \) of \( \theta^* \), Javanmard and Montanari (2014) adds a correction term to \( \hat{\theta} \) to remove the bias introduced by the regularization. A key step in their procedure lies in searching for an approximation \( M \) of the inverse of the population Hessian \( \left[ \mathbb{E} \left( X_i^T X_i \right) \right]^{-1} \) and establishing a finite-sample bound on \( \left| \frac{M}{n} \sum_{i=1}^n X_i^T X_i - I \right|_\infty \) where \( | \cdot |_\infty \) is the elementwise \( l_\infty \)-norm and \( I \) is the identity matrix.

This approximation step becomes more involved in the analysis of a probit-like model \( \mathbb{E} (Y_i | X_i) = \Phi (X_i\theta^*) \) as the inverse of the population Hessian depends on the \( p \)-dimensional coefficient vector \( \theta^* \) (where \( p \) is allowed to exceed \( n \)) and consequently its approximation, \( M \), will inevitably rely on the initial Lasso estimate \( \hat{\theta} \) of \( \theta^* \). Letting \( \hat{H}(\hat{\theta}) \) denote the sample Hessian of our interest evaluated at \( \hat{\theta} \), we use a discretization argument and a metric entropy result to derive a finite-sample bound on \( \left| M \hat{H}(\hat{\theta}) - I \right|_\infty \). In contrast to Javanmard and Montanari (2014) where the asymptotic normality is derived by conditioning on the covariates \( X \), our results incorporate the additional randomness in both \( X \) and \( \hat{\theta} \) as our \( M \) also depends on \( \hat{\theta} \) (clearly in the case of linear models as in Javanmard and Montanari, \( M \) would only depend on \( X \)). It turns out doing so requires imposing a growth condition on the number of non-zero components in \( \hat{\theta} \), \( |J(\hat{\theta})| \).

It is possible to trade off the condition on \( |J(\hat{\theta})| \) with a sparsity assumption on the inverse of the population Hessian by applying a different debiasing method proposed by van de Geer, Bühlmann, Ritov, and Dezeure, (2014) as well as Zhang and Zhang (2014). We choose not to adopt their
method as the sparsity assumption on the inverse of the population Hessian does not seem natural to the applications underlying this paper. However, like the analysis in our paper, some of the results in van de Geer, et al. (2014) do not require conditioning on the Xs and can be extended to certain nonlinear probability models such as the fractional probit. This motivates us to compare the conditions in our results with those in Van de Geer, et al. To do so, we abstract our analysis from the panel data setting and make the comparisons specific to problems concerning cross-sectional data and inference about a single parameter.

We apply one of the debiasing methods to estimate the effects of spending on math pass rates for fourth graders in Michigan between 1995 and 2001. Our proposed model includes the full set of Chamberlain’s regressors and the interactions between all variables and year dummies. This specification is much more general and flexible than the fractional probit model proposed in Papke and Wooldridge (2008) which only includes the classical Mundlak time averages to control for the correlated random effects. Our results show that spending has a positive and statistically significant average partial effect on pass rates. The finding on spending is consistent with the story in Papke and Wooldridge (2008) which uses the pooled quasi-maximum likelihood procedure. In terms of the magnitudes, the estimates for the effect of spending based on the debiasing method are also comparable to those in Papke and Wooldridge (2008).

In terms of organization, we first introduce a set of notations that are commonly used in this paper. We then present in Section 2 formal statements of the correlated random effects probit and fractional probit models. The inference procedure is introduced in Section 3. We present the theoretical guarantees of the inference procedure in Section 4 where we also compare our results with those in Javanmard and Montanari (2014) and van de Geer, et al. (2014). In Section 5 we apply the procedure from Section 3 to estimate the effects of spending on math pass rates. Section 6 provides directions for future research and concludes the paper. The proofs are collected in Sections A and B.

Notation. The \( l_q \)-norm of a \( p \)-dimensional vector \( \Delta \) is denoted by \( |\Delta|_q, 1 \leq q \leq \infty \) where \( |\Delta|_q := \left( \sum_{j=1}^{p} |\Delta_j|^q \right)^{1/q} \) when \( 1 \leq q < \infty \) and \( |\Delta|_{\infty} := \max_{j=1,\ldots,p} |\Delta_j| \) when \( q = \infty \). For a matrix \( H \in \mathbb{R}^{p \times p} \), write \( |H|_{\infty} := \max_{i,j} |H_{ij}| \) to be the elementwise \( l_{\infty} \)-norm of \( H \). For a square matrix \( H \), denote its minimum eigenvalue by \( \lambda_{\min}(H) \). For a vector \( \Delta \in \mathbb{R}^p \), let \( J(\Delta) = \{ j \in \{1,\ldots,p\} \mid \Delta_j \neq 0 \} \) be its support, i.e., the set of indices corresponding to its non-zero components \( \Delta_j \). The cardinality of a set \( J \subseteq \{1,\ldots,p\} \) is denoted by \( |J| \) and if \( J = \emptyset \), \( |J| = 0 \). For a set of indices \( A \subseteq \{1,\ldots,p\} \) with cardinality \( |A| \), denote \( \Delta^A \) or \( \Delta_A \) the sub-columns or sub-rows of a vector \( \Delta \) formed by the indices in \( A \). Also, let \( H_j \) denote the \( j \)-th row of \( H \) and \( H_{ij} \) the \( i \)-th column of \( H \). Moreover, define \( |\Delta|_0 = \sum_{j=1}^{p} \mathbb{I}\{\Delta_j \neq 0\} \). For functions \( f(n) \) and \( g(n) \), write \( f(n) \gtrless g(n) \) to mean that \( f(n) \geq cg(n) \) for a universal constant \( c \in (0, \infty) \) and similarly, \( f(n) \lesssim g(n) \) to mean that \( f(n) \leq c'g(n) \) for a universal constant \( c' \in (0, \infty) \), and \( f(n) \asymp g(n) \) when \( f(n) \gtrsim g(n) \) and \( f(n) \lesssim g(n) \) hold simultaneously. Also denote \( \max\{a, b\} \) by \( a \vee b \) and \( \min\{a, b\} \) by \( a \wedge b \). The standard normal c.d.f. and p.d.f. are denoted by \( \Phi(\cdot) \) and \( \phi(\cdot) \), respectively. Also, as a general rule for this paper, all the \( c \) constants denote universal positive constants that are independent of \( n \). The specific values of these constants may change from place to place.

2 Correlated random effects probit

It is helpful to begin with a description of the CRE probit model. For simplicity, we assume a balanced panel data set in this paper. In general, elements of the vector \( W_{it} \) in (1) can be both time constant and time varying, but for now it is notationally convenient to consider only time-varying covariates. In the following, we provide a list of assumptions that are used for the CRE probit model.
Under the normality and strict exogeneity assumptions, classical finite dimensionality assumptions, the heterogeneity to be correlated with time-varying covariates and results in an equation like (5).

Assumption 2.1: (i) The draws \((Y_i, W_i)_{i=1}^n\) are independently and identically distributed. (ii) \(\alpha_i = V_i \gamma_0 + \eta_i\), where \(V_i\) is a set of functions of \(W_i\) (including a constant term) and \(\eta_i|W_i \sim N(0, \sigma^2_\eta)\). (iii) \(\nu_{it}|W_{it}, \alpha_i \sim N(0, 1)\) for all \(t = 1, ..., T\). (iv) \(E(Y_{it}|W_i, \alpha_i) = E(Y_{it}|W_{it}, \alpha_i)\).

Missing from Assumption 2.1 is a clear statement of identification conditions. Those will become clear in Section 4 after pooled estimation of a probit model in high-dimensional settings has been treated.

Given (1) and Assumption 2.1, we can write
\[
Y_{it} = 1 \{X_{it}\theta_0 + u_{it} > 0\},
\]
where \(X_{it} = [W_{it}, V_i]\) and \(u_{it} = \eta_i + v_{it}\), and the \(p\)-dimensional coefficient vector is \(\theta_0 = [\beta^T_0, \gamma^T_0]^T\).

Under the normality and strict exogeneity assumptions,
\[
P(Y_{it} = 1|W_i) = P(\eta_i + v_{it} > -W_{it}\beta_0 - V_i\gamma_0|W_i) = \Phi(W_{it}\beta^* + V_i\gamma^*) = \Phi(X_{it}\theta^*),
\]
where \(\theta^* = \frac{\theta_0}{\sqrt{1+\sigma^2_\eta}}\) is a set of scaled parameters. Note that \(\theta^*\) is the vector we can hope to estimate; we cannot identify \(\theta_0\) and \(\sigma^2_\eta\) separately. However, as discussed in Wooldridge (2010, Chapter 15) in the standard setting with a small number of regressors, \(\theta^*\) is the quantity that indexes the APEs.

For applications with fractional response variables \(Y_{it} \in [0, 1]\), we only use the conditional mean assumption
\[
E(Y_{it}|W_i) = \Phi(X_{it}\theta^*), \ t = 1, ..., T.
\]
Therefore, our analysis applies to the so-called fractional probit model with unobserved heterogeneity proposed by Papke and Wooldridge (2008). That paper uses a standard CRE approach to allow the heterogeneity to be correlated with time-varying covariates and results in an equation like (5).

Thus, the current paper extends the fractional probit model to high dimensional settings which allows \(p\), the dimension of \(\theta^*\), to exceed the sample size \(n\). In regard to applications concerning panel data with fixed number of periods considered in this paper, the high dimensionality of \(\theta^*\) arises from \(\gamma^*\). Allowing the dimension of \(\gamma^*\) to grow with and possibly exceed \(n\) provides us the flexibility to include a large number of approximating terms in \(V_i\). On the other hand, we point out that the method and theory provided in this paper can be extended to cases where a large number of time-varying functions of \(W_{it}\)s are included.

3 Inference procedure

3.1 Pooled quasi-maximum likelihood estimation

A convenient and computationally simple method for estimating the parameters and APEs in CRE probit models – whether applied to binary or fractional outcomes – is pooled quasi-maximum likelihood estimation (QMLE). (The “quasi” is in the case where \(Y_{it}\) is not binary.) In fact, pooled QMLE can be used in a variety of panel data settings, even if we have no interest in an underlying CRE. In particular, we can start with the conditional expectation that has a probit form as in (5).

With fixed \(T\), we can obtain a consistent and \(\sqrt{n}\)-asymptotically normal estimator of \(\theta^*\) in the classical (finite \(p\)) settings, along with various partial effects, by solving the problem
\[
\max_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{t=1}^{T} \{ (1 - Y_{it}) \log [1 - \Phi (X_{it}\theta)] + Y_{it} \log [\Phi (X_{it}\theta)] \},
\]
where $\sum_{t=1}^{T} \{(1 - Y_{it}) \log [1 - \Phi (X_{it}\theta)] + Y_{it} \log [\Phi (X_{it}\theta)]\}$ is the pooled (or partial) quasi-log likelihood function for cross-sectional unit $i$. Wooldridge (2010, Section 13.8) discusses estimation and inference using pooled MLEs and QMLEs generally. Papke and Wooldridge (2008) discuss why solving this problem produces a consistent estimator under (5) only (and mild regularity conditions).

Because the estimator is obtained by treating the data as if it is one long cross section, computation is typically straightforward. Of course, in general the time series dependence in the data must be accounting for when estimating asymptotic variances, which are then used in test statistics and confidence intervals. A general “sandwich” estimator is available, and this estimator accounts for both serial dependence of unknown form and the fact that the probit log likelihood is only a quasi-log likelihood in the fractional response case. See Wooldridge (2010, Section 13.8) for a general discussion. A joint MLE, which requires obtaining the joint distribution of $(Y_{i1}, ..., Y_{iT})$ conditional on $(X_{i1}, ..., X_{iT})$, can be computationally much more difficult, and we may not even have enough assumptions to obtain the joint distribution. In any case, the joint MLE will generally require more assumptions to consistently estimate the parameters (and the benefit from the additional assumptions and computational burden is more asymptotic efficiency). A joint MLE for fractional responses seems especially challenging in cases of practical interest. Therefore, we focus on the pooled (Q)MLE.

As we argue in the next subsection, pooled QMLE can be modified to allow cases with very high dimensional parameters provided we add a penalty.

### 3.2 L1-regularized pooled QMLE and inference

Let $\hat{\theta}$ be a first-step estimator of the $p$-dimensional coefficient vector $\theta^* = \frac{\theta_0}{\sqrt{1 + \sigma_n^2}}$. In this paper, we choose the first-step estimator to be the $l_1$-regularized conditional quasi-maximum likelihood estimator

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \{(1 - Y_{it}) \log [1 - \Phi (X_{it}\theta)] + Y_{it} \log [\Phi (X_{it}\theta)]\} + \lambda_n |\theta|_1 \right\},$$

where $\lambda_n \geq 0$ is a regularization parameter.

Let us consider a finite set $A_1 \subseteq \{1, ..., p\}$ of indices associated with $\theta^*$, where the cardinality of $A_1$ is denoted by $|A_1|$. For example, the set $A_1$ may be a singleton corresponding to some treatment “effect”. Let $A = J(\hat{\theta}) \cup A_1$ and denote the sample “Hessian” associated with $A$ by

$$\hat{H}^A = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} H_{it}^A(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} f(X_{it}\hat{\theta})X_{it}^A X_{it}^A$$

where $f(u) = u^2 \frac{\phi(u)}{\Phi(u)} + \left( \frac{\phi(u)}{\Phi(u)} \right)^2$. Let the score

$$s_i^A(\hat{\theta}) = \sum_{t=1}^{T} s_{it}^A(\hat{\theta}) = -\sum_{t=1}^{T} \frac{\phi(X_{it}\hat{\theta})X_{it}^A (Y_{it} - \Phi(X_{it}\hat{\theta}))}{\Phi(X_{it}\hat{\theta}) (1 - \Phi(X_{it}\hat{\theta}))}$$

for each $i$. Let $M^A$ be a solution that satisfies the following feasibility problem:

$$\left| M^A \hat{H}^A - I^A \right|_{\infty} \leq \mu_n$$

6
where $\mu_n$ is a non-negative tuning parameter and $I^A$ is the $|A| \times |A|$ identity matrix. There are many computationally efficient algorithms for solving programs like this, for example, the interior point method (see e.g., Bertsimas and Tsitsiklis, 1997; Boyd and Vandenberghe, 2004).

Motivated by the idea of “debiasing” in Javanmard and Montanari (2014) used for inference in sparse linear regression models, we define the second-step estimator $\hat{\theta}^A$ as follows:

$$\hat{\theta}^A = \hat{\theta}^A + \frac{M^A}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} s_{it}^A(\hat{\theta}).$$

An estimator of the APE with respect to the (continuous) $j$th covariate ($j \in \{1, \ldots, p\}$) can be obtained by $\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_j f \left( X_{it} \hat{\theta} \right)$ where $\hat{\theta} = \left( \hat{\theta}^A, 0^A^c \right)$ with $A = J(\hat{\theta}) \cup A_1$ and $\hat{\theta}^A$ is obtained from (11). If the $j$th covariate is a binary variable indicating treatment status, then the estimator becomes $\frac{1}{n} \sum_{i=1}^{n} \left[ \Phi \left( X_{it}^{(1)} \hat{\theta} \right) - \Phi \left( X_{it}^{(0)} \hat{\theta} \right) \right]$ where $\Phi \left( X_{it}^{(1)} \hat{\theta} \right)$ is evaluated at $X_{itj} = 1$ and $\Phi \left( X_{it}^{(0)} \hat{\theta} \right)$ is evaluated at $X_{itj} = 0$.

Some intuitions

Let us introduce the short notations $s_i^A(\theta^*) = \sum_{t=1}^{T} s_{it}^A(\theta^*)$ and $H_i^A(\theta) = \sum_{t=1}^{T} H_{it}^A(\theta)$, defined in a similar way as $\sum_{t=1}^{T} s_{it}^A(\hat{\theta})$ in (9) and $\sum_{t=1}^{T} H_{it}^A(\theta)$ in (8), respectively; in particular, we let $H_i^{A*} = \sum_{t=1}^{T} H_{it}^A(\theta^*)$, $\hat{H}_i^A = \sum_{t=1}^{T} H_{it}^A(\hat{\theta})$, and $\hat{H}_i^A = \sum_{t=1}^{T} H_{it}^A(\theta)$. Suppose $\mathbb{E}H_i^{A*}$ and $\mathbb{E}\hat{H}_i^A$ have full rank. We first consider a simple estimator based on (11) in the classical finite $p$ settings where we simply choose $A = \{1, \ldots, p\}$. If $\hat{H}_i^A$ is invertible, then we can set $M^A = (\hat{H}_i^A)^{-1}$ and (11) is simply a Newton-Raphson iteration in the so-called “One-Step Theorem” (Newey and McFadden, 1994), which is used to gain asymptotic efficiency.

More generally, we can search for an $M^A$ such that (10) is satisfied for some “small” positive $\mu_n$ and $M^A \approx (\mathbb{E}\hat{H}_i^A)^{-1}$. To see how $\hat{\theta}^A$ allows us to obtain the asymptotic normality for a finite set $A_1 \subseteq \{1, \ldots, p\}$ of elements in $\hat{\theta}^A$, note that a mean-value expansion of $s_{it}^A(\hat{\theta})$ around $\theta^*$ and some simple algebraic manipulations of (11) yield that

$$\begin{align*}
\hat{\theta}^A - \theta^* & = \frac{M^A}{n} \sum_{i=1}^{n} s_{it}^A(\theta^*) + \left[ I^A - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^A \right] \left( \hat{\theta}^A - \theta^* \right) - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^{AA^c} \theta^* \hat{A}^c \\
& = \left( \mathbb{E}H_i^{A*} \right)^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} s_{it}^A(\theta^*) \right] + \left[ I^A - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^A \right] \left( \hat{\theta}^A - \theta^* \right) - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^{AA^c} \theta^* \hat{A}^c \\
& + \left[ \left( \mathbb{E}\hat{H}_i^A \right)^{-1} - \left( \mathbb{E}H_i^{A*} \right)^{-1} \right] \frac{1}{n} \sum_{i=1}^{n} s_{it}^A(\theta^*) \\
& + \left[ I^A - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^A \right] \left( \hat{\theta}^A - \theta^* \right) - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^{AA^c} \theta^* \hat{A}^c \\
& + \left[ \left( \mathbb{E}\hat{H}_i^A \right)^{-1} - \left( \mathbb{E}H_i^{A*} \right)^{-1} \right] \frac{1}{n} \sum_{i=1}^{n} s_{it}^A(\theta^*) + \left[ I^A - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^A \right] \left( \hat{\theta}^A - \theta^* \right) - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^{AA^c} \theta^* \hat{A}^c \\
& + \left[ \left( \mathbb{E}\hat{H}_i^A \right)^{-1} - \left( \mathbb{E}H_i^{A*} \right)^{-1} \right] \frac{1}{n} \sum_{i=1}^{n} s_{it}^A(\theta^*), \tag{12}
\end{align*}$$

where $\left[ \hat{H}_i^{AA^c} \right]_{j} = \left[ \sum_{t=1}^{T} f(X_{it} \bar{\theta}(j)) X_{it}^A X_{it}^{A^c} \right]_{j}$, and $\bar{\theta}(j) (j \in A)$ is some intermediate value between $\hat{\theta}$ and $\theta^*$ (possibly differing across the rows of the Hessian matrix).

For any finite set $A_1$ of indices in $A$, if we can show that as $n \to \infty$: \(a\) There exists a solution $M^A$ to (10) such that the $A_1$-subrows of (13) multiplied by $\sqrt{n}$ are of smaller order than those of (12) multiplied by $\sqrt{n}$, and \(b\) the $A_1$-subrows of (14) multiplied by $\sqrt{n}$ are of smaller order than those of (12) multiplied by $\sqrt{n}$, then, $\sqrt{n} \left( \hat{\theta}^A_1 - \theta^* A_1 \right)$ has the same asymptotic distribution
as the $A_1$-subrows of (12) multiplied by $\sqrt{n}$. Note that even if $\hat{\theta}^{A_1} = 0$, the elements in $\tilde{\theta}^{A_1}$ are non-zero by (11). Consequently, our procedure allows us to obtain the asymptotic distribution of $\tilde{\theta}^{A_1}$ and conduct inference on functions of $\theta^{*A_1}$.

Establishing the asymptotic normality for the estimator of the APEs is more delicate as it concerns deriving limiting distributions for sample averages of linear functions of a potentially large set of estimates for primary parameters (associated with policy-related variables) as well as nuisance parameters (associated with approximating terms).

4 Theoretical results

In this section, we establish the asymptotic normality of a finite set of components in the second-step estimator obtained by (11), as well as the estimator for the APE with respect to the $j$th covariate ($j \in \{1, ..., p\}$). For notational simplicity, in the theoretical results presented below, we assume the regime of interest is $p \geq n$ or $p \asymp n$. The modification to allow $p \ll n$ is trivial. Before stating the main theorems, we first present the asymptotic bounds for $|\hat{\theta} - \theta^*|_2$ and $|\hat{\theta} - \theta^*|_1$ and illustrate the role of these bounds in obtaining our inference results. To focus on the main point of this paper, we assume that $|\theta^*|_1 \leq 1$ in the analysis to avoid unnecessary complications. For a random variable $V$, we denote its sub-Exponential norm by $|V|_{\varphi_1} := \sup_{r \geq 1} r^{-\frac{1}{2}} (E|V|^r)^{\frac{1}{2}}$ and its sub-Gaussian norm by $|V|_{\Psi_1} := \sup_{r \geq 1} r^{-\frac{1}{2}} (E|V|^r)^{\frac{1}{2}}$. For the following lemma, we also define $\tilde{\sigma} = \max_{j=1,...,p} \sup_{t=1} |X_{it}j|_{\varphi_1}$, $S^\tau := \{j \in \{1, ..., p\} : |\theta_j^*| > \tau\}$, and $\bar{B}_n = \sqrt{\frac{k \log p}{n}} + \left(\frac{|\theta^*_{S^\tau}|_1 \sqrt{\frac{\log p}{n}}}{\tau} \right)^{\frac{1}{2}}$ where $k = |S^\tau|$.

Lemma 4.1: Suppose Assumption 2.1 holds and the random matrix $X_t$ consists of bounded components for all $t = 1, ..., T$. Assume $\lambda_{\min}(\Sigma_{X_t}) \geq \kappa_L \gtrsim 1$ where $\Sigma_{X_t} = E[X_t^T X_t]$. If $\theta^*$ is exactly sparse with at most $k$ non-zero coefficients such that $k = O \left(\frac{n}{\log p}\right)$, we let $\tau = 0$; otherwise we let $\tau = \frac{\lambda_n}{\kappa_L}$. Suppose $|\theta^*|_1 \sqrt{\frac{\log p}{n}} = O(1)$ when $\tau \neq 0$. If $\hat{\theta}$ solves program (7) with $\lambda_n = c_0 \tilde{\sigma} \sqrt{\frac{\log p}{n}}$ for some universal constant $c_0 > 0$, then,

$$\left|\hat{\theta} - \theta^*\right|_2 = O_p \left(\bar{B}_n\right),$$

$$\left|\hat{\theta} - \theta^*\right|_1 = O_p \left(\sqrt{k} \bar{B}_n + |\theta^*_{S^\tau}|_1\right).$$

By the definition of a sub-Gaussian random variable (e.g., Vershynin, 2012), boundedness of $X_t$ guarantees that $\tilde{\sigma} \gtrsim 1$. Recalling the second term in (13), bound (16) plays an important role in ensuring that the $A_1$-subrows of this term multiplied by $\sqrt{n}$ is negligible asymptotically relative to the $A_1$-subrows of the leading term $\left[\mathbb{E} H^{*A}\right]^{-1} \sum_{i=1}^{n} H_i^A(\theta^*)$ in (12) multiplied by $\sqrt{n}$. To see this, note that

$$\left|\left[I^A - M^A \sum_{i=1}^{n} H_i^A\right] (\hat{\theta}^A - \theta^{*A})\right|_\infty \leq \left|I^A - M^A \sum_{i=1}^{n} H_i^A\right|_\infty \left|\hat{\theta}^A - \theta^{*A}\right|_1.$$

4.1 Asymptotic normality

We now proceed with the results on the asymptotic normality. The following definitions and assumptions are needed for the next two theorems.
Definitions. Let the set
\[ S_{r_1, r_2} := \{ \theta \in \mathbb{R}^p \mid |\theta - \theta^*_1|_1 \leq r_1, |\theta - \theta^*_2|_2 \leq r_2 \}, \]
where \( r_1 = c_2 \left[ \sqrt{k} \bar{B}_n + |\theta^*_2|_1 \right] \), the upper bound in (16), and \( r_2 = c_1 \bar{B}_n \), the upper bound in (15), where \( c_1 \) and \( c_2 \) are some sufficiently large universal positive constants. Also, for a given set of indices \( A \subseteq \{1, ..., p\} \) and \( j, j' \in A \), define
\[
T^A_{1,j,j'}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbb{E} H^A(\theta)^{-1} X^{TA}_{it} X^A_{it} \right]_{jj'},
\]
\[
T^A_{2,j,j'}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbb{E} H^A(\theta)^{-1} \sum_{t=1}^{T} f'(X_{it}\theta) X^{TA}_{it} X^A_{it} \right] \left[ \mathbb{E} H^A(\theta)^{-1} X^{TA}_{it} X^A_{it} \right]_{jj'},
\]
where \( \mathbb{E} H^A(\theta) := \mathbb{E} \left( H^A(\theta) \right) \), \( f(u) = u \frac{\phi(u)}{\phi(\psi)} + \left( \frac{\phi(u)}{\phi(\psi)} \right) \) and \( f'(u) \) denotes the first derivative of \( f \) evaluated at \( u \), and \([H]_{jj'}\) denotes the \( jj'\) th entry of a matrix \( H \).

Assumption 4.2 (Local Identification): Given the set of indices \( A \subseteq \{1, ..., p\} \), for some parameter \( \kappa > 0 \), \( \lambda_{\min} \left( \mathbb{E} H^A(\theta) \right) \geq \kappa \) for all \( \theta \in S_{r_1, r_2} \).

Remark. In terms of the set of indices \( A \subseteq \{1, ..., p\} \), Assumption 4.2 requires that the population 'Hessian' associated with \( A \) and evaluated at any \( \theta \) belonging to a local neighborhood \( S_{r_1, r_2} \) of \( \theta^* \) have full rank. By the definition of a sub-Exponential random variable (e.g., Vershynin, 2012), boundedness of \( X_t \) together with Assumption 4.2 ensures that there exist parameters \( \sigma_{1A}, \sigma_{2A}, \sigma_{3A}, \sigma_{4A} > 0 \) such that for all \( \theta \in S_{r_1, r_2} \),
\[
\sigma_{1A} \geq \max_{j, j' \in A} \left[ \sum_{t=1}^{T} \left| \mathbb{E} H^A(\theta)^{-1} X^{TA}_{it} X^A_{it} \right|_\psi \right]_{jj'}, \tag{17}
\]
\[
\sigma_{2A} \geq \max_{j \in A} \left[ \sum_{t=1}^{T} \left| \mathbb{E} H^A(\theta)^{-1} X^{TA}_{it} \right|_\psi \right]_{jj'}, \tag{18}
\]
\[
\sigma_{3A} \geq \max_{j \in A, j' \in \{1, ..., p\}} \left[ \sum_{t=1}^{T} \left| \mathbb{E} H^A(\theta)^{-1} X^{TA}_{it} X^A_{it} \right|_\psi \right]_{jj'}, \tag{19}
\]
\[
\sigma_{4A} \geq \max_{j \in A, j' \in \{1, ..., p\}} \left[ \sum_{t=1}^{T} \left| \mathbb{E} H^A(\theta)^{-1} X^{TA}_{it} X^A_{it} \right|_\psi \right]_{jj'} \tag{20}.
\]

Assumption 4.3: Recalling the set \( S_T^\Delta \) defined for Lemma 4.1, \( \frac{\Delta T \Sigma X_t \Delta}{|\Delta|^2} \leq \kappa_U < \infty \) for all \( \Delta \in \mathbb{R}^p \setminus \{0\} : |\Delta S_T^\Delta|_1 \leq 3|\Delta S_T^\Delta|_1 + 4|\theta^*_2 S_T|_1 \) and \( t = 1, ..., T; \) for any fixed unit vector \( \Delta, |X_{it} \Delta|_\psi \leq \sigma_{X_t} \).

Our first inference result concerns the asymptotic normality of a finite set of components in
the second-step estimator \( \hat{\theta}^A \) obtained by (11). For notational convenience, we denote \( E_1 = \sqrt{\sigma_{3A}} \vee \max_{j:j'} \sqrt{\mathbb{E} \left[ \sup_{\theta \in \mathcal{S}_{r_1, r_2}} T_{1, j:j'}^A(\theta) \right]} \), \( E_2 = \sqrt{\sigma_{4A}} \vee \max_{j:j'} \sqrt{\mathbb{E} \left[ \sup_{\theta \in \mathcal{S}_{r_1, r_2}} T_{2, j:j'}^A(\theta) \right]} \), and

\[
C_t^* = \sum_{t=1}^T \left( \max \left\{ \kappa_U, \sigma_{X_t}, E_2^2 \right\} + \max \left\{ \kappa_U, \sigma_{X_t}, E_2^2 \right\} \right).
\]

**Theorem 4.1.** Given the set of indices \( A = J(\hat{\theta}) \cup A_1 \) for a finite set \( A_1 \) of our interests, suppose:

(i) Assumptions 4.1-4.3 and the conditions in Lemma 4.1 hold; (ii) \( \left( \theta^{*}_{A_1} \right)_{1} \rightarrow \frac{\log p}{n^{2+\varsigma}} \) for some constant \( \varsigma > 0 \), and \( \frac{\log p}{n^{2+\varsigma}} = o(1) \); (iii) \( C_t^*(k^{1/2}) \frac{\log p}{\sqrt{\bar{n}}} = o(1) \) and \( \sigma_{1A} \frac{\log p}{n} \ll \bar{B}_n \); (iv) we choose \( \mu_n = c\sigma_{1A} \frac{\log p}{n} \) in (10) for some universal constant \( c > 0 \); (v) the cardinality of \( A \), \( |A| = o_p \left( \frac{n}{(kV1)\log p} \right) \). Then, for the second-step estimator \( \tilde{\theta}^A \) in (11), we have

\[
\sqrt{n} \left( \tilde{\theta}^A - \theta^{*}_A \right) \overset{d}{\to} N \left( 0, \Sigma^A \right)
\]

where \( \tilde{\theta}^{A_1} \) denotes the \( A_1 \)-subcomponent of \( \tilde{\theta}^A \) and \( \Sigma^A \) denotes the \( |A| \times |A| \) sub-matrix of

\[
\Sigma^A := \text{var} \left( \left[ \mathbb{E} H^{*A} \right]^{-1} \sum_{t=1}^T s^A_{it}(\theta^*) \right).
\]

Note that the partial debiasing step (11) (a Newton-Raphson like\(^1\) iteration) yields the same asymptotic distribution as a post-Lasso approach where inference is based upon the estimates returned by (6) with the regressors in \( A = J(\hat{\theta}) \cup A_1 \). The condition \( \left| \theta^{*}_{A_1} \right|_1 \geq \frac{\log p}{n^{2+\varsigma}} \) in Theorem 4.1 ensures that the \( A_1 \)-subrows of the last term \( -\frac{M^A}{n} \sum_{i=1}^n \bar{H}_i^{A_1}\theta^{*A_1} \) in (13) multiplied by \( \sqrt{n} \) is negligible asymptotically relative to the \( A_1 \)-subrows of the leading term \( \left[ \mathbb{E} H^{*A} \right]^{-1} \frac{M^A}{n} \sum_{i=1}^n s^A_{i}(\theta^*) \) in (12) multiplied by \( \sqrt{n} \). The condition \( \left| \theta^{*}_{A_1} \right|_1 \geq \frac{\log p}{n^{2+\varsigma}} \) is needed in the proof that determines the scale of \( \mu_n \) (see Lemma A1).

We now turn to our main result which establishes the asymptotic normality of the estimator for the APE with respect to the (continuous) \( j \)th covariate. A similar result can be established for the APE estimator \( \frac{1}{n} \sum_{i=1}^n \left[ \Phi \left( X^{(1)}_{it}\hat{\theta} \right) - \Phi \left( X^{(0)}_{it}\hat{\theta} \right) \right] \) if the \( j \)th covariate is a binary variable indicating treatment status, where \( \Phi \left( X_{it}\hat{\theta} \right) \) is evaluated at \( X_{it} = 1 \) and \( \Phi \left( X_{it}\hat{\theta} \right) \) is evaluated at \( X_{it} = 0 \).

For Theorem 4.2, we denote

\[
\xi_i^A := \left( \left[ \mathbb{E} H^{*A} \right]^{-1} \sum_{t=1}^T s^A_{it}, \left( \left[ \mathbb{E} H^{*A} \right]^{-1} \sum_{t=1}^T s^A_{it} \right)^T \right)^T,
\]

\[
a := \left( \mathbb{E} \left[ \phi(X_{it}\theta^*) \right], \mathbb{E} \left[ \theta^*_j \phi(X_{it}\theta^*)X^A_{it} \right] \right)^T.
\]

**Theorem 4.2.** For a finite set \( A_1 \) of our interests, let \( A = J(\hat{\theta}) \cup A_1 \) and \( \hat{\vartheta} = (\hat{\theta}^A, 0^{A^c}) \) with \( \hat{\theta}^A \) obtained from debiasing \( \hat{\theta}^A \) in (11). Suppose \( \left( \theta^{*}_{A_1} \right)_{1} \rightarrow \frac{\log p}{n^{2+\varsigma}} \) for some constant \( \varsigma > 0 \).

\(^1\)The “like” here means we are taking a Newton-Raphson iteration based on \( H^A = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T f(X_{it}\hat{\theta})X^{TA}_{it}X^A_{it} \) instead of the \( |A| \times |A| \) submatrix of \( \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T f(X_{it}\hat{\theta})X^{TA}_{it}X^A_{it} \) associated with the set \( A \).
Also assume \( |\theta^*|^2 \sqrt{\frac{\log p}{n}} = o(1) \), \( \sigma_1 A \sqrt{\frac{\log p}{n}} \gtrsim \tilde{B}_n \), \( |A| = o_p \left( \left( \frac{n}{(k \vee 1) \log p} \right)^{\frac{1}{4}} \right) \), and the following conditions hold,
\[
\frac{|a_{j1}|^2 |A|^2}{\text{var} \left( \sum_{i=1}^{n} a_i^T \xi_i^A \right)} \log \left( \frac{|a_{j1}|^2 |A|^2 \log n}{\text{var} (a_i^T \xi_i^A)} \right) = o_p(1), \tag{19}
\]
\[
n \left[ \text{var} \left( \sum_{i=1}^{n} a_i^T \xi_i^A \right) \right]^{-\frac{1}{2}} \left( \sqrt{\frac{\log p}{n^{1+\kappa}}} \sqrt{C_1^*(k \vee 1)^2} \log p \right) = o_p(1), \tag{20}
\]
where \( C_1^* \) is defined in Theorem 4.1. For the estimator of the APE with respect to the (continuous) \( j \)th covariate \( (j \in \{1, \ldots, p\}) \) and \( t \in \{1, \ldots, T\} \), under Assumptions (i) and (iv) of Theorem 4.1, we have
\[
\frac{1}{\sqrt{\text{var} \left( \sum_{i=1}^{n} a_i^T \xi_i^A \right)}} \sum_{i=1}^{n} \left[ \hat{\theta}_j \phi \left( X_{it} \hat{\theta} \right) - \theta_{jt}^* \phi \left( X_{it} \theta^* \right) \right] \overset{d}{\to} \mathcal{N}(0, 1). \tag{21}
\]

Condition (19) imposes restrictions on the growth rate of \( |J(\hat{\theta})| \) (i.e., the number of non-zero coefficients in \( \hat{\theta} \)) since \( A = J(\hat{\theta}) \cup A_1 \) and \( A_1 \) is a finite set. Let us consider a special case where \( \text{var} (a_i^T \xi_i^A) \propto \sum_{t=1}^{T} \text{var} \left( a_i^T \xi_i^A \right) \propto |A| \) and \( |a_{j1}|^2 \propto |A|^2 \). Then the pre-multiplier \( \frac{1}{\sqrt{\text{var} \left( \sum_{i=1}^{n} a_i^T \xi_i^A \right)}} \)
required for asymptotic normality in Theorem 4.2 is of order no greater than \( \frac{1}{\sqrt{n}} \) and condition (20) implies that \( \frac{\sqrt{\log p}}{n^{1/4} \sqrt{|J(\hat{\theta})| \vee 1}} \sqrt{C_1^*(k \vee 1)^2} \log p = o_p(1) \). Moreover, (19) implies that
\[
|J(\hat{\theta})|^3 \left( \log |J(\hat{\theta})| + \log \log n \right) = O_p(n),
\]
which is satisfied under the condition \( |A| = o_p \left( \left( \frac{n}{(k \vee 1) \log p} \right)^{\frac{1}{4}} \right) \).

For Theorem 4.2, note that the restriction, \( |A| = o_p \left( \left( \frac{n}{(k \vee 1) \log p} \right)^{\frac{1}{4}} \right) \), is stronger than the one, \( |A| = o_p \left( \frac{n}{(k \vee 1) \log p} \right) \), in Theorem 4.1. This is not surprising as the APEs involve linear combinations of the coefficient vector \( \theta^* \). If the parameter \( k \) (defined prior to Lemma 4.1) is finite and \( |J(\hat{\theta})| = O_p(k \vee 1) \) with high probability\(^2\), then \( \sqrt{n} \)-asymptotic normality is achieved by our APE estimator.

### 4.2 Estimators of the asymptotic variances

A “sandwich” estimator for \( \Sigma^{A_1} \) in (18) is \( \left[ \hat{\Sigma}^A \right]^{A_1} = \left[ M^A \hat{B}^A M^A^T \right]^{A_1} \), where \( M^A \) is obtained from (10) and \( \hat{B}^A = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} s_{it}^A(\hat{\theta}) s_{it}^A(\hat{\theta})^T \). An estimator of \( \mathbb{E} \left[ \left( a_i^T \xi_i^A \right)^2 \right] \) in (21) is
\[
\hat{\Lambda}_t^j = \frac{1}{n} \sum_{i=1}^{n} (a_i^T \xi_i^A)^2,
\]
where
\[
\hat{\xi}_i^A := \left( M^A \sum_{t=1}^{T} s_{it}^A(\hat{\theta}), M^A \sum_{t=1}^{T} s_{it}^A(\hat{\theta})^T \right)^T,
\]
\[
\hat{\alpha} := \left( \frac{1}{n} \sum_{i=1}^{n} \phi(X_{it} \hat{\theta}), \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_j \phi(X_{it} \hat{\theta}) X_{it}^A \right)^T.
\]
\(^2\)Indeed, \( |J(\hat{\theta})| = O_p(k \vee 1) \) can be ensured under a “sparse eigenvalue condition” (Bickel, et. al, 2009; Belloni and Chernozhukov, 2013).
Provided that \( \frac{(k\nu 1) \log p}{n} = o(1) \), the consistency of \( \left[ \hat{\Sigma}^A \right]^{A_1} \) (and consequently \( \hat{\Lambda}_i^A \)) requires
\[
|J(\hat{\theta})| = O_p \left( \left( \frac{n}{(k\nu 1) \log p} \right)^{\frac{1}{2} - \varepsilon} \right) \quad (\text{where } \varepsilon \in (0, \frac{1}{4})) \] as each component in the matrix \( \hat{\Sigma}^A = M^A \hat{B}^A M^{AT} \) is a sum of approximately \( \left( |J(\hat{\theta})| \right)^2 \) terms with each term converging to its population counterpart at the rate \( \sqrt{\frac{(k\nu 1) \log p}{n}} \). Note that the restrictions on \( |J(\hat{\theta})| \) required for consistency of the covariance estimator and asymptotic normality of our APE estimator in Theorem 4.2 coincide.

If we assume that the score (evaluated at \( \theta^* \)) is serially uncorrelated, i.e., \( \mathbb{E} \left[ s_i^A s_i^{sAT} \right] = 0 \) for all \( t \neq t' \), then \( \mathbb{E} \left[ \sum_{t=1}^{T} H_i^A (\theta^*) \right] = -\mathbb{E} \left[ \sum_{t=1}^{T} \sum_{t'=1}^{T} s_{it}^A (\theta^*) s_{it'}^A (\theta^*)^T \right] \) and consequently \( \Sigma^A = \left( \mathbb{E} \left[ \sum_{t=1}^{T} H_i^A (\theta^*) \right] \right)^{-1} \) which can be estimated consistently by \( M^A \) without any further restriction on \( |J(\hat{\theta})| \) (i.e., Assumption (v) in Theorem 4.1 would suffice); on the other hand, to obtain a consistent estimator for \( \mathbb{E} \left[ \left( a^T \xi A \right)^2 \right] \), we still require \( |J(\hat{\theta})| = o_p \left( \left( \frac{n}{(k\nu 1) \log p} \right)^{\frac{1}{2}} \right) \) as \( \hat{\Lambda}_i^A = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{a}_i^T \hat{\xi}_i^A \right)^2 \) is a sum of approximately \( \left( |J(\hat{\theta})| \right)^2 \) terms with each term converging to its population counterpart at the rate \( \sqrt{\frac{(k\nu 1) \log p}{n}} \).

As a side comment, if the unknown function of interest belongs to the \( \alpha \)th order Sobolev class and a classical series estimation procedure is adopted, it appears that the condition \( \alpha > 2 \) is needed. To see this, recall the optimal rate of convergence for a series type estimator for these classes of functions is \( \left( \frac{1}{n} \right)^{\frac{\alpha}{2\alpha + 1}} \) and the corresponding number of basis terms included to achieve such a rate is \( n^{\frac{1}{2\alpha + 1}} \). In order for the covariance estimators to be consistent, we need \( \left( \frac{1}{n} \right)^{\frac{\alpha}{2\alpha + 1}} n^{\frac{2}{2\alpha + 1}} = o(1) \), which holds when \( \alpha > 2 \).

### 4.3 Comparison with the linear case in Javanmard and Montanari (2014)

In what follows, we abstract our analysis from the panel data setting and make comparisons with Javanmard and Montanari (2014) by looking at problems concerning cross-sectional data and inference about a single parameter. For the case of a sparse linear regression \( Y_i = X_i \theta^* + u_i \) with cross-sectional data, Javanmard and Montanari let \( A = \{1, ..., p\} \) in (10) and debias all components of \( \hat{\theta} \) in (11), which leads to the following expansion
\[
\hat{\theta} - \theta^* = \frac{M}{n} \sum_{i=1}^{n} X_i^T (Y_i - X_i \theta^*) + \left[ I - \frac{M}{n} \sum_{i=1}^{n} X_i^T X_i \right] (\hat{\theta} - \theta^*).
\]

Consequently, as long as the asymptotic normality is derived by conditioning on \( X \) (which is the case for Javanmard and Montanari, 2014), Assumptions (ii) and (v) in Theorem 4.1 can be dropped.

In the current setting, it makes little sense condition on the sequence of covariates, as we are explicitly allowing correlation between the heterogeneity and covariates. Therefore, we incorporate the additional randomness in both \( X \) and \( \hat{\theta} \) (note that our \( M \) would also depend on \( \hat{\theta} \) for probit-like models and clearly for linear models considered in Javanmard and Montanari, \( M \) would only depend on \( X \)). In our context, we have to impose a restriction on the number of non-zero coefficients in the Lasso – \( |J(\hat{\theta})| = o_p \left( \sqrt{\frac{n}{(k\nu 1) \log p}} \right) \). Such a restriction is essentially necessary for our analysis by noting that each row in (13)-(14) involves multiplications of vectors whose dimensions depend on \( |J(\hat{\theta})| \). It is possible to trade off the condition \( |J(\hat{\theta})| = \)}
\[ o_p \left( \sqrt{\frac{n}{(k+1) \log p}} \right) \]
with another assumption by applying a different method for approximating the inverse of the Hessian, as we discuss in the subsequent section.

### 4.4 Comparison with the debiasing procedure via nodewise Lasso

van de Geer, Bühlmann, Ritov, and Dezeure (2014) as well as Zhang and Zhang (2014) propose an alternative procedure for approximating the inverse of the Hessian by applying the Lasso \( p \) times for each regression problem \( X_j \) versus \( X_{-j} \) (the design submatrix without the \( j \)th column). We decided to adopt the method by Javanmard and Montanari (2014) as it appears more natural for panel data applications concerning probit functions. Later in the course of this research, we realize that Van de Geer, et al. (2014) also contain inference results regarding a single parameter in Generalized Linear Models with cross-sectional data; moreover, like the analysis in our paper, some of their results do not require conditioning on the \( X \)s (while the conditioning argument is used in Javanmard and Montanari 2014). Therefore, comparing the conditions in our results (Theorem 4.1) with those in Van de Geer, et al. (2014) would be both interesting and worthwhile. To do so, we also abstract our analysis from the panel data setting and make the comparisons specific to problems concerning cross-sectional data and inference about a single parameter.

First, our results do not require the inverse of the population Hessian matrix to be sparse, which is imposed in van de Geer, et al. (2014). On the other hand, we do have to restrict the growth rate on \( |J(\hat{\theta})| \). To be precise, van de Geer, et al. assumes that the off diagonal entries in the inverse of the Hessian are \( s \)-sparse where \( |s| = o \left( \sqrt{\frac{n}{\log p}} \right) \); note that this assumption can be related to an \( s \)-sparse node regression of \( X_j \) versus \( X_{-j} \) for every \( j = 1, \ldots, p \). Under such a condition, they show that \( \left| M_j - H^*_j \right|_1 = o_p(1) \) where \( M_j \) is the \( j \)th row of their inverse Hessian approximation (see Corollary 3.1 in Van de Geer, et al.). To arrive at a similar result without imposing the sparsity condition on the off diagonal entries of \( (H^*)^{-1} \), we instead assume \( |J(\hat{\theta})| = o_p \left( \sqrt{\frac{n}{(k+1) \log p}} \right) \) (i.e., Assumption (v) in Theorem 4.1); note that this assumption can be related to a “sparse eigenvalue condition” in literature (e.g., Belloni and Chernozhukov, 2013).

We are not confident whether the sparsity structure for the inverse of the Hessian can be a reasonable assumption for economic applications underlying this paper. However, there are theoretical benefits for imposing such a condition and using the nodewise Lasso to obtain the inverse Hessian approximation: it is useful for obtaining a limit theory that has certain uniform properties (e.g., a local uniformity at least); in addition, besides Assumption (v), Assumption (ii) in Theorem 4.1 can also be dropped.

We must emphasize that even with a sparsity condition on the inverse of the Hessian, applying the debiasing procedure via nodewise Lasso would not help rid of Assumption (ii) in Theorem 4.2 for the results on the APE estimators (except for the case where the dimension of \( \theta^* \) increases with \( n \) at a sufficiently small rate). This observation highlights the theoretical difficulty in the inference of APEs (as opposed to a single or low dimensional parameters), which essentially arise from the fact that establishing the asymptotic normality for the APE estimators concerns deriving limiting distributions for sample averages of linear functions of a potentially large set of estimates for primary parameters as well as nuisance parameters when a series approximation for the conditional mean of the unobserved heterogeneity is considered.

Nevertheless, the additional \( s \)-sparsity condition (with \( |s| = o \left( \sqrt{\frac{n}{\log p}} \right) \)) on the inverse of the Hessian and the debiasing procedure via nodewise Lasso can help weaken the restriction on \( |J(\hat{\theta})| \)
in Theorem 4.2. In particular, we can replace the condition $|J(\hat{\theta})| = o_p\left(\frac{n}{(k^2+1)\log p}\right)$ with $|J(\hat{\theta})| = o_p\left(\frac{n}{(k^2+1)\log p}\right)$.

5 Application to test pass rates

We apply the debiasing method to estimate the effects of spending on math test pass rates for fourth graders in Michigan after funding for schools was changed from a local, property-tax based system to a statewide system supported primarily through a higher sales tax (and lottery profits) in 1994. We use the district-level data from Papke and Wooldridge (2008), which includes 501 school districts for the years 1995 through 2001. A detailed description of this data set is provided in Papke and Wooldridge (2008).

The response variable, $math4$, is the fraction of fourth graders passing the Michigan Education Assessment Program (MEAP) fourth-grade math test in the district. The explanatory variables in our model include the same spending measure used in Papke and Wooldridge (in logarithmic form, $\log(\text{avgrexp})$), district enrollment ($\log(\text{enroll})$), the fraction of students eligible for the free and reduced-price lunch programs ($\text{lunch}$, a proxy for poverty levels), year dummies from 1996 to 2001 ($\text{year}$), and the interactions between the year dummies and the three time-varying variables. These explanatory variables correspond to the time-varying regressors, $W$, in (4). The policy variable is $\log(\text{avgrexp})$; the other variables are included as controls. To control for the random effects that are correlated with $\log(\text{avgrexp})$, $\log(\text{enroll})$, and $\text{lunch}$, we include their time averages and the Chamberlain’s regressors from 1996 to 2001. These explanatory variables correspond to the time-constant controls, $V$, in (4).

To summarize, our specification for $X_{it}\theta^*$ in (5) takes on the following form

$$X_{it}\theta^* = \alpha + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \log(\text{enroll}_{it}) + \beta_3 \text{lunch}_{it} + \delta_t + \pi_1 \log(\text{avgrexp}_{it}) + \pi_2 \log(\text{enroll}_{it}) + \pi_3 \text{lunch}_{it} + \gamma_1 \log(\text{avgrexp}_{i,96}) + \gamma_2 \log(\text{enroll}_{i,96}) + \gamma_3 \text{lunch}_{i,96} + \gamma_4 \log(\text{avgrexp}_{i,01}) + \gamma_5 \log(\text{enroll}_{i,01}) + \gamma_6 \text{lunch}_{i,01},$$

which amounts to a total of $p = 49$ regressors including the intercept, while $n = 501$ is the number of cross sectional observations used to estimate (22).

To implement the debiasing method via (10)-(11), we consider

$$A = J(\hat{\beta}_2, \hat{\beta}_3, \hat{\delta}, \hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\gamma}) \cup A_1$$

(23)

where $A_1$ is the set of indices corresponding to $\{\hat{\alpha}, \hat{\beta}_1\}$. For comparison purposes, we also consider

$$B = J(\hat{\delta}, \hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\gamma}) \cup B_1$$

(24)

where $B_1$ is the set of indices corresponding to $\{\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$. In applications like the current one, which is essentially a policy analysis, one can make a case for forcing the policy variable or variables into the parameter vector (whether they are selected by the Lasso or not) at the debiasing step as in (23) or (24). In (24), we also force $\log(\text{enroll})$ and $\text{lunch}$ into the parameter vector at the
debiasing step. In any case, our interest is squarely on the average partial effect of \( \log(\text{avgrexp}) \) and not on the effects of the other variables.

We consider several \( \lambda_n \geq 0.55 \sqrt{\frac{\log p}{n}} \) in (7). For \( \lambda_n \) in this range and the choice of \( A \) in (23) or \( B \) in (24), it suffices to set \( \mu_n = 0 \) in (10) and consequently \( M^S = \left( \hat{H}^S \right)^{-1} \) (\( S = A \) or \( S = B \)) in (11) where \( \hat{H}^S \) is the sample Hessian defined in (8). In addition to the debiasing method, we consider another high-dimensional procedure, the so-called Post Lasso, which follows the same first-step Taylor series expansion to the first order condition of the loss function in (6).

\[ \hat{\beta} = \left( \hat{H}^S \right)^{-1} \hat{H}_n \hat{H}_n' \hat{H}^S \hat{H}_n' \hat{\theta} \]

\[ \tilde{\beta}_1 = \hat{\beta}_1 - \frac{1}{n} \hat{H}_n \hat{H}_n' \hat{\theta} \]

\[ \tilde{\beta}_2 = \hat{\beta}_2 - \frac{1}{n} \hat{H}_n \hat{H}_n' \hat{\theta} \]

\[ \tilde{\beta}_3 = \hat{\beta}_3 - \frac{1}{n} \hat{H}_n \hat{H}_n' \hat{\theta} \]

Simple evocation of the asymptotic normality theory from the classical finite \( \hat{\beta} \) settings by applying a Taylor series expansion to the first order condition of the loss function in (6).

Table 1 reports the first-step estimates obtained from (7) under three choices of \( \lambda_{nj} = c_j \sqrt{\frac{\log p}{n}} \) where \( c_1 = 0.55, c_2 = 0.6, \) and \( c_3 = 0.65 \). Variables that are not reported in Table 1 simply have coefficients estimated to be 0 by the Lasso. Tables 2-3 report the second-step parameter estimate \( \tilde{\beta}_1 \) and the APE estimate \( (\text{APE}) \) of \( \log(\text{avgrexp}) \), along with the standard errors \( (se) \), given by the debiasing method and the Post Lasso method, under \( \lambda_{nj}, j = 1, 2, 3, \) respectively. In particular, Table 2 is based on \( A \) in (23) and Table 3 is based on \( B \) in (24).

To compare the high-dimensional methods with the classical inference procedure, we apply (6) to estimate the following two specifications where \( X_{it}\theta^* \) in (5) takes on the forms

\[ X_{it}\theta^* = \alpha + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \log(\text{enroll}_{it}) + \beta_3 \text{lunch}_{it} + \delta_t, \]

\[ X_{it}\theta^* = \alpha + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \log(\text{enroll}_{it}) + \beta_3 \text{lunch}_{it} + \delta_t + \gamma_1 \log(\text{avgrexp}_i) + \gamma_2 \log(\text{enroll}_i) + \gamma_3 \text{lunch}_i. \]

The results via the classical methods are reported in Table 4 where Columns “Classical” are for (25) and Columns “Classical Mundlak” are for (26). The APE estimates and their standard errors are computed for each \( t = 95, ..., 01 \); the reported \( A\text{PE} \) and standard errors in Tables 2-4 are simply the time averages.

The results regarding spending in Tables 2-3 agree with the finding in Papke and Wooldridge (2008) which applies the pooled QMLE to estimate (26) (“Classical Mundlak” in Table 4); namely, spending has a positive and statistically significant average partial effect on math pass rates. A ten percent increase in average spending increases the pass rate by about three percentage points, on average. In terms of the magnitudes, overall, the estimates for the effect of spending based on the high-dimensional methods are more comparable to those based on “Classical Mundlak” than to those based on “Classical”. Under \( \lambda_{n2} \) and \( \lambda_{n3} \), the debiasing method and the “Classical Mundlak” method yield very similar estimates. As the penalty level increases, the first-step parameter estimate associated with spending in Table 1, the corresponding debiased estimate and the APE estimate in Tables 2-3 decrease; in contrast, the Post Lasso estimates increase as the penalty level increases. Under \( \lambda_{n3} \), the debiasing method and the Post Lasso method yield almost identical estimates.

<table>
<thead>
<tr>
<th>( \log(\text{avgrexp}) )</th>
<th>( \text{year}_{97} )</th>
<th>( \text{year}_{99} )</th>
<th>( \text{lunch}_{99} )</th>
<th>( \text{lunch}_{01} )</th>
<th>( \text{year}_{96} \cdot \text{lunch} )</th>
<th>( \text{year}_{97} \cdot \text{lunch} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_{n1} )</td>
<td>0.152</td>
<td>-0.013</td>
<td>0.001</td>
<td>-0.238</td>
<td>-0.355</td>
<td>-0.081</td>
</tr>
<tr>
<td>( \lambda_{n2} )</td>
<td>0.105</td>
<td>0</td>
<td>0</td>
<td>-0.211</td>
<td>-0.341</td>
<td>-0.029</td>
</tr>
<tr>
<td>( \lambda_{n3} )</td>
<td>0.056</td>
<td>0</td>
<td>0</td>
<td>-0.185</td>
<td>-0.328</td>
<td>0</td>
</tr>
</tbody>
</table>

| \( \lambda_{n1} \) | 0.152          | -0.013         | 0.001          | -0.238         | -0.355         | -0.081         | -0.236         |
| \( \lambda_{n2} \) | 0.105          | 0              | 0              | -0.211         | -0.341         | -0.029         | -0.213         |
| \( \lambda_{n3} \) | 0.056          | 0              | 0              | -0.185         | -0.328         | 0              | -0.159         |

Table 1: First-step parameter estimates from the Lasso
would change at least as a function of the number of time periods, and possibly the other covariates. The reason to allow a heteroskedastic function is that the variance of the heterogeneity one should allow for a heteroskedastic-probit function rather than just a standard probit response estimation is used on unbalanced panels. Again, conceptually the extension is straightforward, as this is very standard in analyzing unbalanced panels, and is implicitly assumed when fixed effects (4), selection can be correlated with unobserved heterogeneity, but not idiosyncratic shocks. Therefore, in the context of equation, how to modify pooled QMLEs to allow selection to be correlated with observed covariates and fractional probit models with panel data where the number of time periods is fixed and small relative to the number of cross-sectional observations. In particular, our procedure allows us to model the conditional mean of the unobserved effect with a much larger number of approximating terms relative to the classical approaches for a probit model. It also extends the pooled quasi-maximum likelihood estimator of the fractional probit model in Papke and Wooldridge (2008) to high dimensional settings. We apply the debiasing method to estimate the effects of spending on test pass rates. Our results show that spending has a positive and statistically significant average partial effect, with magnitudes similar to those found by Papke and Wooldridge (2008).

It is natural to think the approach here can be extended to other nonlinear models. Conceptually there is no issue, as pooled quasi-maximum likelihood is applicable to many kinds of response variables, including count variables and nonnegative continuous variables. It may be possible to verify the regularity conditions when Yit is unbounded, so that pooled Poisson estimation of CRE models can be used. One can also argue that allowing for extensions of the probit response function, such as a heteroskedastic probit model, would be worth pursuing.

Our approach is easily extended to unbalanced panels, provided we assume that selection is appropriately ignorable. In the context of fully parametric CRE models, Wooldridge (2016) shows how to modify pooled QMLES to allow selection to be correlated with observed covariates and unobserved heterogeneity, but not idiosyncratic shocks. Therefore, in the context of equation (4), selection can be correlated with \( W_{it} : t = 1, ..., T \) and \( \eta_i \), but not with \( \nu_{it} : t = 1, ..., T \). This is very standard in analyzing unbalanced panels, and is implicitly assumed when fixed effects estimation is used on unbalanced panels. Again, conceptually the extension is straightforward, as one should allow for a heteroskedastic-probit function rather than just a standard probit response function. The reason to allow a heteroskedasticity function is that the variance of the heterogeneity would change at least as a function of the number of time periods, and possibly the other covariates

<table>
<thead>
<tr>
<th>( \lambda_{n1} )</th>
<th>Post Lasso</th>
<th>( \hat{\beta}_1 )</th>
<th>se(( \hat{\beta}_1 ))</th>
<th>( \hat{APE} )</th>
<th>se(( \hat{APE} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.634</td>
<td>0.083</td>
<td>0.215</td>
<td>0.028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.683</td>
<td>0.084</td>
<td>0.232</td>
<td>0.028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.778</td>
<td>0.086</td>
<td>0.265</td>
<td>0.020</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Second-step estimates and standard errors for spending, with \( A \)

<table>
<thead>
<tr>
<th>( \lambda_{n1} )</th>
<th>Post Lasso</th>
<th>( \hat{\beta}_1 )</th>
<th>se(( \hat{\beta}_1 ))</th>
<th>( \hat{APE} )</th>
<th>se(( \hat{APE} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.653</td>
<td>0.087</td>
<td>0.221</td>
<td>0.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.703</td>
<td>0.089</td>
<td>0.238</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.797</td>
<td>0.091</td>
<td>0.271</td>
<td>0.031</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Second-step estimates and standard errors for spending, with \( B \)

<table>
<thead>
<tr>
<th>Classical Mundlak</th>
<th>Classical</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>se(( \hat{\beta}_1 ))</td>
</tr>
<tr>
<td>0.333</td>
<td>0.094</td>
</tr>
</tbody>
</table>

Table 4: Estimates and standard errors for spending

6 Conclusion and future directions

We have proposed a simple method for inference of the average partial effects (APEs) in probit and fractional probit models with panel data where the number of time periods is fixed and small relative to the number of cross-sectional observations. In particular, our procedure allows us to model the conditional mean of the unobserved effect with a much larger number of approximating terms relative to the classical approaches for a probit model. It also extends the pooled quasi-maximum likelihood estimator of the fractional probit model in Papke and Wooldridge (2008) to high dimensional settings. We apply the debiasing method to estimate the effects of spending on test pass rates. Our results show that spending has a positive and statistically significant average partial effect, with magnitudes similar to those found by Papke and Wooldridge (2008).

It is natural to think the approach here can be extended to other nonlinear models. Conceptually there is no issue, as pooled quasi-maximum likelihood is applicable to many kinds of response variables, including count variables and nonnegative continuous variables. It may be possible to verify the regularity conditions when \( Y_{it} \) is unbounded, so that pooled Poisson estimation of CRE models can be used. One can also argue that allowing for extensions of the probit response function, such as a heteroskedastic probit model, would be worth pursuing.

Our approach is easily extended to unbalanced panels, provided we assume that selection is appropriately ignorable. In the context of fully parametric CRE models, Wooldridge (2016) shows how to modify pooled QMLES to allow selection to be correlated with observed covariates and unobserved heterogeneity, but not idiosyncratic shocks. Therefore, in the context of equation (4), selection can be correlated with \( \{ W_{it} : t = 1, ..., T \} \) and \( \eta_i \), but not with \( \{ \nu_{it} : t = 1, ..., T \} \). This is very standard in analyzing unbalanced panels, and is implicitly assumed when fixed effects estimation is used on unbalanced panels. Again, conceptually the extension is straightforward, as one should allow for a heteroskedastic-probit function rather than just a standard probit response function. The reason to allow a heteroskedasticity function is that the variance of the heterogeneity would change at least as a function of the number of time periods, and possibly the other covariates...
as well. Now, we would take $V_i$ to be functions of

$$\{(S_{it}, S_{it}W_{it}) : t = 1, \ldots, T\},$$

where $S_{it}$ is a binary selection indicator that determines whether we observe a complete case for unit $i$ in time period $t$. Functions include the time averages of covariates of the complete cases and also functions of the selection indicators themselves, such as the number of time periods observed for unit $i$, say $T_i$. Once the $V_i$ have been chosen, the objective function in equation (7) is simply multiplied by the selection indicator, $S_{it}$, and the probit response function is replaced with probit with heteroskedasticity. Estimation is somewhat more challenging but quite feasible. Without the penalty it would be pooled heteroskedastic probit estimation where the variance function depends on a relatively small number of elements of $V_i$ – see Wooldridge (2016). The challenge is deriving the asymptotic properties of the estimator with additional nonlinearity in the estimation, under high dimensionality and sparsity.

A Main proofs

Proof of Lemma 4.1. We can write (7) as

$$L_n(\theta) = \left\{ \frac{-1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ (1 - Y_{it}) \log \left[ 1 - \Phi (X_{it} \theta) \right] + Y_{it} \log \left[ \Phi (X_{it} \theta) \right] \right\} + \lambda_n |\theta|_1 \right\},$$

Define the following quantity

$$\delta L_n(\theta^*, \Delta) := L_n(\theta^* + \Delta) - L_n(\theta^*) - \left\langle \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} s_{it}(\theta^*), \Delta \right\rangle.$$

To prove Lemma 4.1, we verify conditions (G1) and (G2) in Negahban, et. al (2010), as well as show the following two steps: Step 1. $\delta L_n(\theta^*, \Delta)$ satisfies the restricted strong convexity (RSC) condition defined in Negahban, et. al (2010), where $\Delta = \hat{\theta} - \theta^*$; Step 2.

$$\left| \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} s_{it}(\theta^*) \right|_\infty \leq \sqrt{\tilde{\sigma}^2 \log \frac{p}{n}},$$

with high probability, where $\tilde{\sigma} = \max_{j=1,\ldots,p} \left| \sum_{t=1}^{T} |X_{itj}|_{\psi_1} \right|$. Then we can apply Theorem 1 in Negahban, et. al (2010) to obtain

$$|\hat{\theta} - \theta^*|_2 \leq \max \left\{ \tilde{\sigma}, c' \frac{\lambda_n \sqrt{k}}{\kappa_L} + c' \frac{\lambda_n \sqrt{\sum_{\ell=1}^{p} |1|}}{\kappa_L} \right\},$$

where $\tilde{\sigma} = c\tilde{B}_n$ and $\lambda_n \geq 2 \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} s_{it}(\theta^*) \right|_\infty$.

Condition (G2) holds trivially since we are applying the $l_1$–regularization. To verify condition (G1), note that $L''_n(\theta)$ is

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ (1 - \var{u_{it} \mid u_{it} \leq X_{it} \theta}) + 1(Y_{it} = 0) [1 - \var{u_{it} \mid u_{it} \geq -X_{it} \theta}] \right\} X_{it}^T X_{it}.$$
Since the unconditional variance is normalized to 1 and truncation always reduces variances (Greene, 2003), the scalar term in the expression above is strictly positive.

**Step 1:** For the case where \( \theta^* \) is exactly sparse with at most \( k \) non-zero coefficients such that \( k \lesssim \frac{n}{\log p} \) and \( \tau = 0 \), under the assumption \( \lambda_{\min}(\Sigma_{X_t}) \geq \kappa_L > 0 \), Proposition 2 along with Lemma 1 in Negahban, et. al (2010) yields the RSC of \( \delta L_n(\theta^*, \hat{\Delta}) \). To show the approximately sparse case where we set \( \tau = \frac{\lambda_n}{\kappa_L} \) for a positive universal constant \( c \), we apply the deviation inequality in Lemma 3 from Negahban, et. al (2010) and show that

\[
0 \geq L_n(\theta^* + \hat{\Delta}) - L_n(\theta^*) + \lambda_n \left\{ |(\theta^*_{S_{\tau}} + \hat{\Delta}_{S_{\tau}}, \theta^*_{S_{\tau}^c} + \hat{\Delta}_{S_{\tau}^c})|_1 - |\theta^*_{S_{\tau}}|_1 - |\theta^*_{S_{\tau}^c}|_1 \right\} \\
\geq \frac{\lambda_n}{2} \left\{ -3|\hat{\Delta}_{S_{\tau}}|_1 + |\hat{\Delta}_{S_{\tau}^c}|_1 - 4|\theta^*_{S_{\tau}^c}|_1 \right\}.
\]

Consequently,

\[
|\hat{\Delta}|_1 \leq 4|\hat{\Delta}_{S_{\tau}}|_1 + 4|\theta^*_{S_{\tau}^c}|_1 \leq 4\sqrt{k} |\hat{\Delta}|_2 + 4|\theta^*_{S_{\tau}^c}|_1,
\]

where \( k = |S_{\tau}| \). For the case where \( \tau \neq 0 \), we upper bound the cardinality of \( k \) in terms of the threshold \( \tau \) and the \( l_1 \)-ball with “radius” of \( R = |\theta^*|_1 \). Note that we have

\[
R = \sum_{j=1}^{p} |\theta^*_j| \geq \sum_{j \in S_{\tau}} |\theta^*_j| \geq k_{\tau}
\]

and therefore \( k \leq \tau^{-1}R \). Then, combining the bound in Proposition 2 in Negahban, et. al (2010) with

\[
|\Delta|_1 \leq 4|\Delta_{S_{\tau}}|_1 + 4|\theta^*_{S_{\tau}^c}|_1 \leq 4\sqrt{k} |\Delta|_2 + 4|\theta^*_{S_{\tau}^c}|_1 \leq 4 \sqrt{\tau^{-1}R} |\Delta|_2 + 4|\theta^*_{S_{\tau}^c}|_1,
\]

yields

\[
\delta L_n(\theta^*, \Delta) \geq |\Delta|_2 \left\{ c_1 \kappa_L - b_0 \sqrt{R_{\tau}^{-1} \log p \over n} \right\} - b_1 |\theta^*_{S_{\tau}^c}|_1 \sqrt{\log p \over n} |\Delta|_2
\]

for some constants \( b_0, b_1 > 0 \). With the choice of \( \delta^* \propto (\kappa_L)^{1 \over 2} |\theta^*_{S_{\tau}^c}|_1 \sqrt{\log p \over n} \), if \( \sqrt{R_{\tau}^{-1} \log p \over n} \lesssim \kappa_L \), we have

\[
\delta L_n(\theta^*, \Delta) \geq \delta^* \kappa_L \left\{ |\Delta|_2 \left| - {\frac{1}{2}} \right| \right\} = \delta^* \kappa_L |\Delta|_2
\]

for any \( \Delta \) such that \( |\Delta|_2 \geq \delta^* \) (if \( |\Delta|_2 < \delta^* \), bound (27) holds trivially). For the case where \( \tau = 0 \), we have

\[
\delta L_n(\theta^*, \Delta) \geq |\Delta|_2 \left\{ c_1 \kappa_L - b_0 \sqrt{\log p \over n} \right\} \geq \delta^* \kappa_L |\Delta|_2
\]

as long as \( b_0 \sqrt{\log p \over n} \lesssim \kappa_L \).

**Step 2:** In the following, we show that \( \left| \frac{1}{n} \sum_{t=1}^{n} \sum_{t=1}^{T} s_{it}^{(\theta^*)} \right|_\infty \leq \sqrt{\sigma^2 \log p \over n} \) with high probability.

Note that

\[
\sum_{t=1}^{T} s_{itj}^{(\theta^*)} = - \sum_{t=1}^{T} \frac{\phi(X_{it}^{\theta^*} X_{itj} (Y_{it} - \Phi(X_{it}^{\theta^*}))}{\Phi(X_{it}^{\theta^*} (1 - \Phi(X_{it}^{\theta^*}))}
\]

Recall that the random matrix \( X_t \) consists of bounded random variables for all \( t = 1, ..., T \) and \( |\theta^*|_1 \leq 1 \). Then, \( \Phi(X_{it}^{\theta^*} (1 - \Phi(X_{it}^{\theta^*}) \leq \bar{c} \) and by the definition of the sub-Gaussian norm of a random
variable $V$, $|V|_{\psi_1} := \sup_{r \geq 1} r^{-\frac{3}{2}} (\mathbb{E} |V|^r)^{\frac{1}{r}}$, we can show that the random variable $\sum_{i=1}^{T} s_{itj}(\theta^*)$ is a zero-mean sub-Gaussian with parameter at most $2\bar{c} \max_{j=1,\ldots,p} \left| \sum_{i=1}^{T} |X_{itj}| \right|_{\psi_1}$. Applying a standard sub-Gaussian tail bound yields

$$\left| \frac{1}{n} \sum_{i=1}^{T} \sum_{t=1}^{n} s_{itj}(\theta^*) \right|_{\infty} \leq c_1 \tilde{\sigma} \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - O\left( \frac{1}{n} \right)$, where $c_1 > 0$ is a universal constant.

Combining Step 1 and Step 2 yields

$$|\hat{\theta} - \theta^*|_2 \leq \max\left\{ \tilde{\sigma}, \frac{c_1 \tilde{\sigma}}{\kappa_L} \sqrt{\frac{k \log p}{n}} + c_1 \left( \frac{\| \theta^*_S \|_1 \tilde{\sigma}}{\kappa_L} \sqrt{\frac{\log p}{n}} \right)^{\frac{1}{2}} \right\} := \max\left\{ \tilde{\sigma}, c_1 \tilde{B}_n \right\}$$

with probability at least $1 - O\left( \frac{1}{n} \right)$, for some universal constant $c_1 > 0$. Setting $\tilde{\sigma}$ in Assumption 4.1 to $\tilde{\sigma} = c_1 \tilde{B}_n$ where $c_1 > c > 0$ yields (15) and applying (28) yields (16). □

**Remark.** In the following, we provide an upper bound on $\frac{|X_i \Delta|^2}{n}$, which will be used in the proof later. The second bound in Lemma B2 implies

$$\frac{|X_i \Delta|^2}{n} \leq \frac{3 \kappa_U}{2} \left| \hat{\Delta} \right|^2 + \frac{\alpha' \log p}{n} \left| \hat{\Delta} \right|^2$$

$$\leq \left( \frac{3 \kappa_U}{2} - c' \alpha' \log p \right) \left| \hat{\Delta} \right|^2 + c'' \alpha' \log p \left| \theta^*_S \right|^2$$

$$\leq \frac{(3 + \varsigma') \kappa_U}{2} \left| \hat{\Delta} \right|^2$$

(30)

for a $\varsigma' > 0$, where $\alpha'$ is a constant depending on $\kappa_L$ and $\kappa_U$, and the last inequality follows from the conditions $\frac{\log p}{n} \log \frac{1}{R} \geq \frac{3 \kappa_U}{2}$ and $\frac{\log p}{n} \left| \theta^*_S \right|^2 = o(1)$.

**Proof of Theorem 4.1.** By the expansion in (12)-(14), it suffices to show that, for any finite set $A_1$ of parameters in $A$, as $n \to \infty$: (a) there exists a solution $M^A$ to (10) such that the $A_1$—subrows of (13) multiplied by $\sqrt{n}$ are of smaller order than those of (12) multiplied by $\sqrt{n}$, and (b) the $A_1$—subrows of (14) multiplied by $\sqrt{n}$ are of smaller order than those of (12) multiplied by $\sqrt{n}$, then, $\sqrt{n} \left( \tilde{\theta}^{A_1} - \theta^* A_1 \right)$ has the same asymptotic distribution as the $A_1$—subrows of (12) multiplied by $\sqrt{n}$.

By Lemma A2, we have

$$\sqrt{n} \left[ I^A - M^A \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{H}_{it}^A \right] (\hat{\theta}^A - \theta^A) = o_p(1).$$

Furthermore, note that

$$\left| \left( \mathbb{E} \hat{H}^A \right)^{-1} \mathbb{E} H^* A - I^A \right|_{\infty} \leq \left| \left( \mathbb{E} \hat{H}^A \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \hat{H}_{i}^A - I^A \right|_{\infty} + \left| \left( \mathbb{E} \hat{H}^A \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} (H_{i}^* A - \hat{H}_{i}^A) \right|_{\infty}$$

$$+ \left| \left( \mathbb{E} \hat{H}^A \right)^{-1} \left( \mathbb{E} H^* A - \frac{1}{n} \sum_{i=1}^{n} H_{i}^* A \right) \right|_{\infty} = O_p \left( \sqrt{\frac{(k+1) \log p}{n}} \right),$$

19
where a similar argument that we use to show (44) is applied to bound the second and third terms above, and the bound on the first term follows from Lemma A1. Consequently if \(|A| = o_p \left( \sqrt{\frac{n}{(k+1) \log p}} \right)\), we apply argument similar to those in Lemma A3 with \(\xi_{ij}^A = \left[ \mathbb{E} H^{*A} \right]^{-1} \sum_{t=1}^T s_{it}^A (\theta^*)\) and obtain

\[
\left[ \left( \mathbb{E} \hat{H}^A \right)^{-1} - \left[ \mathbb{E} H^{*A} \right]^{-1} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{t=1}^T s_{it}^A (\theta^*) \right]^{A_1} = o_p(1).
\]

Using the fact established above and that \(|M^n (\mathbb{E} \hat{H}^A) - I^A|_\infty = o_p\left( \sqrt{\frac{\log p}{n}} \right)\) by (45), we obtain

\[
\left[ \left( M^A - \mathbb{E} \hat{H}^A \right)^{-1} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{t=1}^T s_{it}^A (\theta^*) \right]^{A_1} = o_p(1)
\]

where the last line follows since \(|A| = o_p \left( \sqrt{\frac{n}{\log p}} \right)\).

For the third term in (13), with the fact that \(\frac{M^n_A}{n} \sum_{i=1}^n \tilde{H}_{it} \theta^* \theta^* \) in Theorem 4.1, we conclude that the term \(\frac{M^n_A}{n} \sum_{i=1}^n \tilde{H}_{it} \theta^* \theta^* = o_p(1)\). Now, applying Lemma A4 and the Cramer-Wold Theorem yields the claim in Theorem 4.1.

**Proof of Theorem 4.2.** By the construction \(\tilde{\theta} = (\tilde{\theta}^A, 0^{A_c})\) with \(A = J(\theta) \cup A_1\) and \(\tilde{\theta}^A\) obtained from debiasing \(\hat{\theta}^A\) in (11), we have

\[
\left[ \text{var} \left( \sum_{i=1}^n a^T \xi_{i}^A \right) \right]^{-\frac{1}{2}} \sum_{i=1}^n \left[ \hat{\theta}_j \phi \left( X_{it} \hat{\theta} \right) - \theta_j^* \phi \left( X_{it} \theta^* \right) \right]
\]

\[
= \left[ \text{var} \left( \sum_{i=1}^n a^T \xi_{i}^A \right) \right]^{-\frac{1}{2}} \left\{ \sum_{i=1}^n \left[ \hat{\theta}_j \phi \left( X_{it} \hat{\theta} \right) - \theta_j^* \phi \left( X_{it} \theta^* \right) \right] + \sum_{i=1}^n \left[ \theta_j^* \phi \left( X_{it} \hat{\theta} \right) - \theta_j^* \phi \left( X_{it} \theta^* \right) \right] \right\}
\]

\[
= \left[ \text{var} \left( \sum_{i=1}^n a^T \xi_{i}^A \right) \right]^{-\frac{1}{2}} \left\{ (\hat{\theta}_j - \theta_j^*) \mathbb{E} \left[ \phi \left( X_{it} \theta^* \right) \right] + (\hat{\theta} - \theta^*)^T \mathbb{E} \left[ \theta_j^* \phi \left( X_{it} \theta^* \right) X_{it}^T \right] \right\} + \text{REM}
\]

Note that (12)-(14) and (31) imply

\[
\left[ \text{var} \left( \sum_{i=1}^n a^T \xi_{i}^A \right) \right]^{-\frac{1}{2}} \sum_{i=1}^n \left[ \hat{\theta}_j \phi \left( X_{it} \hat{\theta} \right) - \theta_j^* \phi \left( X_{it} \theta^* \right) \right]
\]

\[
= \left[ \text{var} \left( \sum_{i=1}^n a^T \xi_{i}^A \right) \right]^{-\frac{1}{2}} \left\{ \mathbb{E} \left[ \phi \left( X_{it} \theta^* \right) \right] \left[ \mathbb{E} H^{*A} \right]^{-1} \sum_{i=1}^n s_i^A (\theta^*) \right\}
\]

\[
+ \left[ \text{var} \left( \sum_{i=1}^n a^T \xi_{i}^A \right) \right]^{-\frac{1}{2}} \left\{ \mathbb{E} \left[ \theta_j^* \phi \left( X_{it} \theta^* \right) X_{it}^T \right] \left[ \mathbb{E} H^{*A} \right]^{-1} \sum_{i=1}^n s_i^A (\theta^*) \right\} + \text{REM}
\]
where
\[
REM = REM_1 + REM_2
\]
\[
+ \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \left[ \frac{1}{n} \sum_{i=1}^{n} \phi(X_i \bar{\theta}) - \mathbb{E} \left[ \phi(X_i \theta^*) \right] \right] \left[ \mathbb{E} H^* A \right]^{-1} \sum_{i=1}^{n} s_i^A(\theta^*)
\]
\[
+ \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \theta_j^* \phi'(X_i \theta') X_i^A - \mathbb{E} \left[ \theta_j^* \phi'(X_i \theta^*) X_i^A \right] \left[ \mathbb{E} H^* A \right]^{-1} \sum_{i=1}^{n} s_i^A(\theta^*)
\]
\[
- n \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \theta^{*T} \left[ \frac{1}{n} \sum_{i=1}^{n} \theta_j^* \phi'(X_i \theta') X_i^{A^T} \right]
\]
and
\[
REM_1 = \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \left[ \frac{1}{n} \sum_{i=1}^{n} \phi(X_i \bar{\theta}) \right] \left[ M^A - \left( \mathbb{E} \hat{H}^A \right)^{-1} \right] \sum_{i=1}^{n} s_i^A(\theta^*)
\]
\[
+ n \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \phi(X_i \bar{\theta}) \left[ I^A - \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i \right] \left( \hat{\theta}^A - \theta^{*A} \right)
\]
\[
+ n \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \phi(X_i \bar{\theta}) \left[ \frac{M^A}{n} \sum_{i=1}^{n} \hat{H}_i^{A^c} \theta^{*A^c} \right] \left( \theta^A - \hat{\theta}^A \right)
\]
\[
+ \left[ \text{var} \left( \sum_{i=1}^{n} a^T \xi_i^A \right) \right]^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \phi(X_i \bar{\theta}) \left[ \left( \mathbb{E} \hat{H}^A \right)^{-1} - \left( \mathbb{E} H^* A \right)^{-1} \right] \sum_{i=1}^{n} s_i^A(\theta^*),
\]
for some intermediate value \( \theta' \) between \( \theta^* \) and \( \bar{\theta} \). By Lemmas A2 and A5 and the argument in the proof for Theorem 4.1, under the conditions in Theorem 4.2, we can show that the term \( REM = o_p(1) \) and therefore \( REM \) is negligible asymptotically relative to the terms in (32) and (33). We now apply Lemma A3 with \( A = \bar{A} = J(\bar{\theta}) \cup A_1 \) and
\[
\xi_i^A := \left( \mathbb{E} \left( H_i^A \right) \right)^{-1} \sum_{t=1}^{T} s_{it}^A, \left( \mathbb{E} \left( H_i^A \right) \right)^{-1} \sum_{t=1}^{T} s_{it}^A)^T
\]
\[
a := \left( \mathbb{E} \left[ \phi(X_i \theta^*) \right], \mathbb{E} \left[ \theta_j^* \phi'(X_i \theta^*) X_i^A \right] \right)^T.
\]
which yields
\[
\frac{1}{\sqrt{\text{var}(\sum_{i=1}^{n} a_i \xi_i^j)}} \sum_{i=1}^{n} a_i^T \xi_i^j \overset{d}{\to} N(0, 1).
\]

**Lemma A1.** Suppose \[\left| \theta_{\Sigma}^c \right|_1 \leq \frac{\sqrt{\log p}}{n^{1/3}}\]. For a set of indices \(A \subseteq \{1, ..., p\}\), define
\[
E_1 = \sqrt{\sigma_{2A}} \max_{j, j'} \sqrt{\mathbb{E} \left[ \sup_{\theta \in \mathcal{S}_{1, r_2}} T_{1, tjj'}^A(\theta) \right]},
\]
\[
E_2 = \sqrt{\sigma_{4A}} \max_{j, j'} \sqrt{\mathbb{E} \left[ \sup_{\theta \in \mathcal{S}_{1, r_2}} T_{2, tjj'}^A(\theta) \right]},
\]
\[
C_j^* = \sum_{t=1}^{T} \left( \left( \theta^*_1 \right) \max \left\{ \kappa_U, \sigma_{X_i}, E_1^2 \right\} + \max \left\{ \kappa_U, \sigma_{X_i}, E_2^2 \right\} \right),
\]
\[
T_{1, tjj'}^A(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( \mathbb{E} H^A(\theta) \right)^{-1} X_{it}^A X_{it}^j \right]_{jj'},
\]
\[
T_{2, tjj'}^A(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( \mathbb{E} H^A(\theta) \right)^{-1} \mathbb{E} \left[ \sum_{t=1}^{T} f'(X_{it} \theta) X_{it}^A X_{it}^j \right] \left[ \mathbb{E} H^A(\theta) \right]^{-1} X_{it}^A X_{it}^j \right]_{jj'},
\]
where \(\mathbb{E} H^A(\theta) := \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} H_i^A(\theta) \right)\). Let \(\mu^* := |I^A - [\mathbb{E} H^A]^{-1} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{H}_i^A|_\infty\) and \(E_1(\theta)\) be the event that \(\mu^* \leq \mu_n = c_1 \sqrt{\log p \over n}\) for a constant \(c > 0\). Then, if \(C_j^* B_n^2 \leq \frac{\sqrt{\sigma_{4A} \log p}}{n}\), we have \(P[E_1(\theta)] \geq 1 - c_1 \exp(-c_2 \log p)\).

**Proof.** Define the set
\[
\mathcal{S}_{r_1, r_2} := \{ \theta \in \mathbb{R}^p \mid |\theta - \theta^*|^1 \leq r_1, |\theta - \theta^*|^2 \leq r_2 \},
\]
where \(r_1 = c_2 \left[ \sqrt{4B_n + \left( \theta_{\Sigma}^c \right)_{1}^2} \right] \), the upper bound in (16), and \(r_2 = c_1 \sqrt{4B_n}\), the upper bound in (15).

Also define the function
\[
\mathcal{I}(Y_{it} = 1) \left[ 1 - \text{var}(u_{it} \mid u_{it} \leq u) \right] + \mathcal{I}(Y_{it} = 0) \left[ 1 - \text{var}(u_{it} \mid u_{it} \geq -u) \right] = 1 - \text{var}(u_{it} \mid u_{it} \leq u) = \frac{u \phi(u)}{\Phi(u)} + \left( \frac{\phi(u)}{\Phi(u)} \right)^2 = f(u),
\]

where the second and third lines follow from the properties of a standard normal distribution.

To show Lemma A1, we show that
\[
\left| \Sigma^A - I^A \right|_{\infty} := \sup_{\theta \in \mathcal{S}_{r_1, r_2}} \left| \left[ \mathbb{E} H^A(\theta) \right]^{-1} H^A(\theta) - I^A \right|_{\infty} \leq c_{1A} \sqrt{\log p \over n}.
\]

We first show that \(\Sigma_{jj}^A\) is a sub-Exponential variable for \(j, j' = 1, ..., p\). In particular, for \(\theta \in \mathcal{S}_{r_1, r_2}\), we write
\[
\sum_{t=1}^{T} f(X_{it} \theta) \left[ \left[ \mathbb{E} \left[ \sum_{t=1}^{T} f(X_{it} \theta) X_{it}^A X_{it}^j \right] \right]^{-1} X_{it}^A X_{it}^j \right]_{jj'} = \sum_{i,j} U_{t}^{A,i}.
\]
Note that
\[
\left| U_{ijj}^{\theta A} \right|_{\psi} = \sup_{r \geq 1} r^{-1} \mathbb{E} \left[ \left| \sum_{i=1}^{T} f(X_{it}\theta) \left( \mathbb{E} \left[ \sum_{i=1}^{T} f(X_{it}\theta)X_{it}^{TA}X_{it}^{A} \right] \right)^{-1} X_{it}^{TA}X_{it}^{A} \right|_{jjj} \right] \leq \sup_{r \geq 1} r^{-1} \mathbb{E} \left[ \left| \sum_{i=1}^{T} \left( \mathbb{E} \left[ \sum_{i=1}^{T} f(X_{it}\theta)X_{it}^{TA}X_{it}^{A} \right] \right)^{-1} X_{it}^{TA}X_{it}^{A} \right|_{jjj} \right] \leq \sigma_{1A},
\]
where the second inequality follows from \( 0 < f(u) < 1 \) for all \( u \in \mathbb{R} \) for probit together and the last inequality follows from (17). Applying Lemma B1 yields
\[
\mathbb{P} \left( \max_{jjj} \left\{ \left| \Sigma_{jjj}^{A} - I_{jjj}^{A} \right| - \mathbb{E} \left[ \left| \Sigma_{jjj}^{A} - I_{jjj}^{A} \right| \right] \right\} \geq u \right) \leq c_{1} \exp \left( -c_{2} r \left( \frac{n}{\sigma_{1A}^{2}} \wedge \frac{u}{\sigma_{1A}} \right) + 2 \log p \right). \tag{34}
\]
Denote \( T_{jjj}^{A} := \left| \Sigma_{jjj}^{A} - I_{jjj}^{A} \right| \) and using a classical symmetrization argument, we obtain
\[
\mathbb{E} T_{jjj}^{A} \leq 2 \mathbb{E} X_{i,\epsilon} \left[ \sup_{\theta \in S_{r_{1}, r_{2}}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_{i} \right| \left( \mathbb{E} \left[ \sum_{i=1}^{T} f(X_{it}\theta)X_{it}^{TA}X_{it}^{A} \right] \right)^{-1} X_{it}^{TA}X_{it}^{A} \right]_{jjj} \]
\[
= 2 \mathbb{E} X_{i,\epsilon} \left[ \sup_{\theta \in S_{r_{1}, r_{2}}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_{i} g_{jjj}^{A}(X_{i}; \theta) \right| \right]
\]
where \( \{\epsilon_{i} \}_{i=1}^{n} \) is an i.i.d. sequence of Rademacher variables, independent of \( \{X_{i} \}_{i=1}^{n} \). For any \( \delta \in (0, 1], \frac{\delta^{2}}{2} \leq \delta \). We construct a \( \frac{\delta^{2}}{2} \)-covering set of \( S_{r_{1}, r_{2}} \) in the \( l_{2} \)-norm with covering number denoted by \( N_{n} \left( \frac{\delta^{2}}{2}; S_{r_{1}, r_{2}} \right) := N \). For any \( \theta \in S_{r_{1}, r_{2}} \), we can find a \( \theta' \) in the covering set such that \( \left| \theta - \theta' \right|_{2} \leq \frac{\delta^{2}}{2} \). Let \( h(u) := \frac{\phi(u)}{\phi(0)} \). Note that \( \epsilon_{i} \in \{-1, 1\} \) for all \( i = 1, ..., n \), \( h(u) \in (0, 1) \) and \( f(u) \in (0, 1) \) for all \( u \in \mathbb{R} \). Without much loss of generality, we consider the case where \( X_{t} \) consists of bounded random variables supported on \([-1, 1]\). Therefore, for \( \theta' \in \left[ \theta, \theta' \right] \), since \( |X_{it}|_{\infty} \leq 1 \), \( X_{it}\bar{\theta} \leq |X_{it}|_{\infty} |\bar{\theta}|_{1} \approx |\theta'\theta|_{1} \) if \( \sqrt{k}B_{n} + \left| \theta'_{\mathcal{S}_{L}} \right|_{1} \approx (|\theta'|_{1} \vee 1) \), and \( h(X_{it}\bar{\theta}) \approx (|\theta'|_{1} \vee 1) \). We have
\[
g_{jjj}^{A}(X_{i}; \theta) - g_{jjj}^{A}(X_{i}; \theta') = \sum_{t=1}^{T} X_{it}(\theta - \theta') \left( h(X_{it}\bar{\theta}) \right) \Gamma_{1,jjj}^{A}(\bar{\theta}, X_{it}\bar{\theta}, \theta - \theta') + \sum_{t=1}^{T} X_{it}(\theta - \theta') \left( h(X_{it}\bar{\theta}) \right) \Gamma_{1,jjj}^{A}(\bar{\theta}, X_{it}\bar{\theta}, \theta - \theta') + \sum_{t=1}^{T} X_{it}(\theta - \theta') \left( f(X_{it}\bar{\theta}) \right) \Gamma_{2,jjj}^{A}(\bar{\theta}, X_{it}\bar{\theta}, \theta - \theta')
\]
for some intermediate value \( \bar{\theta} \) between \( \theta \) and \( \theta' \), where
\[
\Gamma_{1,jjj}^{A}(\bar{\theta}) = \left[ \mathbb{E} H^{A}(\bar{\theta}) \right]^{-1} X_{it}^{TA}X_{it}^{A},
\]
\[
\Gamma_{2,jjj}^{A}(\bar{\theta}) = \left[ \mathbb{E} H^{A}(\bar{\theta}) \right]^{-1} \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} f(X_{it}\bar{\theta})X_{it}^{TA}X_{it}^{A} \right) \mathbb{E} H^{A}(\bar{\theta})^{-1} X_{it}^{TA}X_{it}^{A}.
\]
We define

\[
T^A_{1,tjj'}(\theta) = \frac{1}{n} \sum_{i=1}^n \left( \left[ \mathbb{E} H^A(\theta) \right]^{-1} X^T_{it} X^A_{it} \right)_{jj'},
\]

\[
T^A_{2,tjj'}(\theta) = \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E} H^A(\theta) \right]^{-1} \mathbb{E} \left[ \sum_{t=1}^T f'(X_{it}\theta) X^T_{it} X^A_{it} \right] \left[ \mathbb{E} H^A(\theta) \right]^{-1} X^T_{it} X^A_{it} \right)_{jj'}.
\]

We now upper bound \( T^A_{1,tjj'}(\bar{\theta}) \) and \( T^A_{2,tjj'}(\bar{\theta}) \). Under Assumption 4.2 and the boundedness of \( X_i \), it can be easily verified that each of the summands in \( T^A_{1,tjj'}(\bar{\theta}) \) and \( T^A_{2,tjj'}(\bar{\theta}) \) is sub-Exponential, respectively. Therefore, we have

\[
P \left( \left| T^A_{1,tjj'}(\bar{\theta}) - \mathbb{E} \left( T^A_{1,tjj'}(\bar{\theta}) \right) \right| > u \right) \leq c_5 \exp \left( -c_6 n \left( \frac{u^2}{\sigma_{3A}^2} \wedge \frac{u}{\sigma_{3A}} \right) \right), \quad (35)
\]

\[
P \left( \left| T^A_{2,tjj'}(\bar{\theta}) - \mathbb{E} \left( T^A_{2,tjj'}(\bar{\theta}) \right) \right| > u \right) \leq c_7 \exp \left( -c_8 n \left( \frac{u^2}{\sigma_{4A}^2} \wedge \frac{u}{\sigma_{4A}} \right) \right). \quad (36)
\]

We also have

\[
\frac{1}{n} \sum_{i=1}^n \epsilon_i g_{jj'}(X_i; \theta) \leq \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{jj'}(X_i; \theta^t) + \frac{1}{n} \sum_{i=1}^n \epsilon_i \left( g_{jj'}(X_i; \theta) - g_{jj'}(X_i; \theta^t) \right) \\
\leq c' \left( \sum_{t=1}^T X_i(\theta - \theta^t) \right) \left[ \left| \theta^t \right|_1 \sqrt{T^A_{1,tjj'}(\bar{\theta})} + \sqrt{T^A_{1,tjj'}(\bar{\theta})} + \sqrt{T^A_{2,tjj'}(\bar{\theta})} \right] \quad (37)
\]

\[
+ \max_{l=1, \ldots, n} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{jj'}(X_i; \theta^t) \right|. \quad (38)
\]

We first upper bound the expectation of the term in (37). Let us fix some \( \tau \geq \max \{ \kappa_U, \sigma_{X_i}, E_1^2 \} \), \( \tau_1 \geq (\kappa_U \vee \sigma_{X_i}) \), and \( \tau_2 \geq E_1^2 \). Let \( \Delta \) be a unit vector. Since \( |X_i\Delta|_n \) and \( \sqrt{T^A_{1,tjj'}(\bar{\theta})} \) (for all \( \theta \in \mathcal{S}_{\tau_1, \tau_2} \)) are both nonnegative, we have

\[
\mathbb{E} \left[ \max_{j,j'} |X_i\Delta|_n \sqrt{T^A_{1,tjj'}(\bar{\theta})} \right] \\
\leq \tau + \int_{\tau}^{\infty} P \left( \max_{j,j'} |X_i\Delta|_n \sqrt{T^A_{1,tjj'}(\bar{\theta})} > v \right) dv \\
\leq \tau + \sum_{j,j'} \int_{\tau}^{\infty} P \left( |X_i\Delta|_n \sqrt{T^A_{1,tjj'}(\bar{\theta})} > v \right) dv \\
\leq \tau + \sum_{j,j'} \int_{\tau_1}^{\infty} P \left( |X_i\Delta|_n^2 > v \right) dv + \sum_{j,j'} \int_{\tau_2}^{\infty} P \left( T^A_{1,tjj'}(\bar{\theta}) > v \right) dv \\
\leq \tau + c' \int_{\tau_1}^{\infty} \exp \left( -c_4 n \left( \frac{v - (\kappa_U \vee \sigma_{X_i})}{\sigma_{X_i}} \right) \right) dv + \int_{\tau_2}^{\infty} \exp \left( -c_6 n \left( \frac{v - E_1^2}{\sigma_{3A}} \right) \right) dv.
\]

The last inequality follows from (35) and

\[
P \left( |X_i\Delta|_n^2 > (\kappa_U \vee \sigma_{X_i}) + u \right) \leq c_3 \exp \left( -c_4 n \left( \frac{u^2}{\sigma_{X_i}^2} \wedge \frac{u}{\sigma_{X_i}} \right) \right). \quad (39)
\]
for any unit vector $\Delta$, where we use the boundedness of $X_i$ and (30). We apply a change of variable $u = v - (\kappa_U \vee \sigma_{X_i})$ and $u = v - E_i^2$, provided $v \geq \tau_1 \geq (\kappa_U \vee \sigma_{X_i})$ and $v \geq \tau_2 \geq E_i^2$, respectively. Integrating yields the following bound

$$
\mathbb{E} \left[ \max_{j, j'} \sup_{\theta \in \mathcal{S}_{r_1, r_2}} \left| X_i^T (\theta - \theta') \right| \sqrt{\frac{T_{1, jj'}^2(\theta)}{n}} \right] \leq c_3 \delta^2 \max \left\{ \kappa_U, \sigma_{X_i}, E_i^2 \right\} + o \left( \frac{\delta^2}{2} \right) \tag{40}
$$

where we use the fact that $\theta'$ is in the $\frac{\delta^2}{2}$-covering set of $\mathcal{S}_{r_1, r_2}$ in the $l_2$-norm. To upper bound the term $\left| X_i(\theta - \theta') \right| \sqrt{T_{2, jj'}^2(\theta)}$, we follow the same argument above with (36) in the place of (35) and obtain

$$
\mathbb{E} \left[ \max_{j, j'} \sup_{\theta \in \mathcal{S}_{r_1, r_2}} \left| X_i(\theta - \theta') \right| \sqrt{T_{2, jj'}^2(\theta)} \right] \leq c_3 \delta^2 \max \left\{ \kappa_U, \sigma_{X_i}, E_i^2 \right\} + o \left( \frac{\delta^2}{2} \right). \tag{41}
$$

We now upper bound the unconditional expectation of the term in (38), the “Rademacher complexity” associated with $T_{jj'}^A$. Conditioning on $X$, it can be easily verified that $rac{1}{n} \sum_{i=1}^n \left( g_{jj'}^A(X_i; \theta') \right)^2 = O(1)$ for all $l = 1, \ldots, N$. Applying the Dudley entropy integral bound as well as the upper bound on $\log N_n(u; \mathcal{S}_{r_1, r_2})$ in Lemma B3, we have, for any $j, j' \in A$,

$$
\mathbb{E} \left[ \max_{i=1, \ldots, N} \left| \sum_{i=1}^n \epsilon_i g_{jj'}^A(X_i; \theta') \right| \right] \lesssim \sqrt{k} \tilde{B}_n + |\theta_{\Sigma \Lambda}^*|_1 \frac{\log p}{n}. \tag{42}
$$

Under the condition $|\theta_{\Sigma \Lambda}^*|_1 \lesssim \sqrt{\frac{\log p}{\nu + c}}$, combining bound (42), (40)-(41) with $\delta^2 \sim \tilde{B}_n^2$, and (34) with $u = c\sigma_A \sqrt{\frac{\log p}{n}}$ yields

$$
\max_{j, j'} \left| \sum_{i=1}^n \tilde{H}_i^A \right| \leq c_0 C_1^* \tilde{B}_n + c_1 \frac{\log p}{n}
$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. Under the condition $C_1^* \tilde{B}_n^2 \lesssim \sigma_A \sqrt{\frac{\log p}{n}}$, the claim in Lemma A1 follows. $\square$

**Lemma A2.** Suppose $|A| = o_p \left( \frac{n}{\log p} \right)$ and $\sigma_A \sqrt{\frac{\log p}{n}} \lesssim \tilde{B}_n$. (a) If $C_1^* \left[ \sqrt{n} \tilde{B}_n \left( \sqrt{k} \tilde{B}_n + |\theta_{\Sigma \Lambda}^*|_1 \right) \right] = o(1)$, then we have

$$
\sqrt{n} \left[ I^A - M^A \frac{1}{n} \sum_{i=1}^n \tilde{H}_i^A \right] (\hat{\theta}^A - \theta^*) = o_p(1).
$$

(b) If $C_1^* \left[ \sqrt{\text{var} \left( \sum_{i=1}^n a_i \xi_i^A \right) \tilde{B}_n \left( \sqrt{k} \tilde{B}_n + |\theta_{\Sigma \Lambda}^*|_1 \right) } \right] = o(1)$, then we have

$$
\frac{1}{\sqrt{\text{var} \left( \sum_{i=1}^n a_i \xi_i^A \right) \tilde{B}_n \left( \sqrt{k} \tilde{B}_n + |\theta_{\Sigma \Lambda}^*|_1 \right) }} \left[ I^A - M^A \frac{1}{n} \sum_{i=1}^n \tilde{H}_i^A \right] (\hat{\theta}^A - \theta^*) = o_p(1),
$$

where $a$ and $\xi_i^A$ are specified in Theorem 4.2.
Proof. We show part (a) below and part (b) follows the same argument where $\sqrt{n}$–scaling is replaced by $n \left[ \text{var}(\sum_{i=1}^{n} a^T \xi^i) \right]^{-\frac{1}{2}}$. Note that we have

$$
\left\| \sqrt{n} \left[ I^A - M^A \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^A \right] (\hat{\theta}^A - \theta^{*A}) \right\|_{\infty}
$$

$$
\leq \sqrt{n} \left| I^A - M^A \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^A \right| \left( \hat{\theta}^A - \theta^{*A} \right) + M^A \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{H}_i^A - \bar{H}_i^A \right) (\hat{\theta}^A - \theta^{*A})
$$

$$
\leq \sqrt{n} \left| I^A - M^A \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^A \right| \left( \hat{\theta}^A - \theta^{*A} \right) + \left( \mathbb{E} \hat{H}^A \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{H}_i^A - \bar{H}_i^A \right) \left| \hat{\theta}^A - \theta^{*A} \right|.
$$

By Lemma A1 and setting $\mu_n = c\sigma_1 A \sqrt{\log p \log n}$ in (10), we are guaranteed that

$$
\left\| M^A \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^A - I^A \right\|_{\infty} \leq c\sigma_1 A \sqrt{\log p \log n} \tag{43}
$$

for any solution $M^A$ satisfying (10). Therefore, by (16) of Lemma 4.1,

$$
\sqrt{n} \left\| M^A \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^A - I^A \right\|_{\infty} \left( \hat{\theta}^A - \theta^{*A} \right) \leq c' \sigma_1 A \sqrt{\log p} \left( \sqrt{k} B_n + \left| \theta^{*s}_{\infty} \right|_1 \right).
$$

It suffices to show that

$$
\left\| \left( \mathbb{E} \hat{H}^A \right)^{-1} \sum_{i=1}^{n} \left( \hat{H}_i^A - \bar{H}_i^A \right) - \mathbb{E} \left[ \left( \hat{H}_i^A - \bar{H}_i^A \right) \right] \right\|_{\infty}.
$$

As a result, we have

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} \left( \hat{H}_i^A - \bar{H}_i^A \right) \right\|_{\infty} \leq cC_2^* \bar{B}_n \tag{44}
$$

with probability at least $1 - O \left( \frac{1}{p} \right)$, where $C_2^* = \sum_{i=1}^{T} \left[ \left( \left( \theta^{*s}_{\infty} \right)_i \right) \mathbb{E} \left[ \left( \theta^{*s}_{\infty} \right)_i \right] \right]$. Applying the bound on $\left\| \hat{\theta}^A - \theta^{*A} \right\|_1$ from Lemma 4.1, under the condition $C_2^* \left( \sqrt{n} B_n + \left| \theta^{*s}_{\infty} \right|_1 \right) = o(1)$, we have

$$
\left\| \left( \mathbb{E} \hat{H}^A \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{H}_i^A - \bar{H}_i^A \right) \right\|_{\infty} \left\| \hat{\theta}^A - \theta^{*A} \right\|_1 = o_p(1).
$$

By bound (43) and the triangle inequality, we have

$$
\left\| \left( M^A \mathbb{E} \hat{H}^A - I^A \right) \left( \mathbb{E} \hat{H}^A \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^A \right\|_{\infty} = O_p \left( \sqrt{\frac{\log p}{n}} \right).
$$
since $|A| = o_p \left( \sqrt{\frac{n}{\log p}} \right)$, we must have

$$|M^A \hat{E} \hat{H}^A - I^A|_\infty = O_p \left( \sqrt{\frac{\log p}{n}} \right)$$

(45)

and therefore, by (44),

$$\left| \left( M^A - \left[ \hat{E} \hat{H}^A \right]^{-1} \right) \frac{1}{n} \sum_{i=1}^{n} \left( \hat{H}^A - \hat{H}^A \right) \right|_\infty = \left| \left( M^A \hat{E} \hat{H}^A - I^A \right) \left[ \hat{E} \hat{H}^A \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{H}^A - \hat{H}^A \right) \right|_\infty$$

$$= o_p \left( C^2 \hat{B}_n \right).$$

As a result,

$$\left| \left( M^A - \left[ \hat{E} \hat{H}^A \right]^{-1} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{H}^A - \hat{H}^A \right) \right|_\infty \left| \hat{\theta}^A - \theta^A \right|_1 = o_p(1).$$

\[\square\]

**Lemma A3.** Let Assumption 2.1 hold. Suppose the random matrix $X_t$ consists of bounded random variables for all $t = 1, \ldots, T$. Given a set of indices $A \subseteq \{1, \ldots, p\}$ with cardinality $|A|$ and a nonrandom matrix $\Sigma$ of dimension $|A| \times |A|$, for a subset $\tilde{A} \subseteq A$, let $\xi_{\tilde{A}} = (\xi^A_{ij})_{j \in \tilde{A}}$ with $\xi^A_{ij} := \Sigma_{j} \sum_{t=1}^{T} s^A_{nit}$. For any $a \in \mathbb{R}^{|\tilde{A}|}$, if

$$\frac{|a|^2 |A|^2}{\operatorname{var}(\sum_{i=1}^{n} a^T \xi^A_{i})} \log \left( \frac{|a|^2 |A|^2 \log n}{\operatorname{var}(a^T \xi^A_{i})} \right) = O_p(1),$$

then for every $\varepsilon > 0$, the Lindeberg condition $\lim_{n \to \infty} \frac{1}{n \rho^4} \sum_{i=1}^{n} \mathbb{E} \left( \left( a^T \xi^A_{i} \right)^2 I \left\{ \left| a^T \xi^A_{i} \right| > \varepsilon \sqrt{n \rho^4} \right\} \right) = 0$ holds almost surely (where $\rho^A_{n} := \frac{1}{n} \sum_{i=1}^{n} \operatorname{var}(a^T \xi^A_{i})$, and $\frac{1}{\sqrt{n \rho^4}} \sum_{i=1}^{n} a^T \xi^A_{i} \overset{d}{\to} \mathcal{N}(0,1)$.

**Proof.** To show the claim in Lemma A3, note that the summands $a^T \xi^A_{i}$ have zero mean and are independent. Let $Y_{it} \sim \Phi(X_{it}\theta^*) = \tilde{Y}_{it}$ and $\Phi(X_{it}\theta^*) \sim \tilde{Y}_{it}$. By boundedness of $X_{it}$ and $|\theta^*| \leq 1$, $b \leq b_{it} \leq \tilde{b}$ for all $i$ and $t$. Also, $\max_{j \in \tilde{A}, t=1,\ldots,T} |\Sigma_j X_{it}^T A| = O (|A|)$. Therefore, $|a^T \xi^A_{i}| \leq c^* |a|^1 \tilde{b} |A|^1 \sum_{t=1}^{T} \tilde{Y}_{it}$ for some constant $c^* > 0$. As a result,

$$\mathbb{E} \left( \left( a^T \xi^A_{i} \right)^2 I \left\{ \left| a^T \xi^A_{i} \right| > \varepsilon \sqrt{n \rho^4} \right\} \right) \leq \mathbb{E} \left( \left( a^T \xi^A_{i} \right)^2 I \left\{ \sum_{t=1}^{T} \tilde{Y}_{it} > c_1 \tilde{b}^{-1} |a|^1 \left( n \rho^4 \right)^{\frac{1}{2}} |A|^{-1} \right\} \right)$$

$$\leq c_2 \left( \sum_{t=1}^{T} \tilde{Y}_{it} \right)^2 I \left\{ \sum_{t=1}^{T} \tilde{Y}_{it} > c_1 \tilde{b}^{-1} |a|^1 \left( n \rho^4 \right)^{\frac{1}{2}} |A|^{-1} \right\}$$

$$\leq c_2 \sqrt{\mathbb{E} \left( \left( \sum_{t=1}^{T} \tilde{Y}_{it} \right)^4 \right)} \sqrt{\mathbb{P} \left( \sum_{t=1}^{T} \tilde{Y}_{it} > c_1 \tilde{b}^{-1} |a|^1 \left( n \rho^4 \right)^{\frac{1}{2}} |A|^{-1} \right)} \left( |a|^2 |A|^2 \tilde{b}^2 \right)$$

$$\leq c_0 |a|^2 |A|^2 \tilde{b}^2 \sqrt{\exp \left( -c_3 \tilde{b}^{-2} |a|^2 \left( n \rho^4 \right)^{2} |A|^{-2} T^{-2} + \log T \right)}$$

27
where we have applied a Cauchy-Schwarz inequality. Consequently, condition (46) implies that the Lindeberg condition holds. An application of the Lindeberg CLT yields $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} a_i^T \xi_i \xrightarrow{d} \mathcal{N}(0, 1)$. □

**Lemma A4.** Suppose Assumption 4.2 and the conditions in Lemma 4.1 hold. Assume $|A| = o_p \left( \sqrt{\frac{n}{\log p}} \right)$. Then $\frac{1}{n} \sum_{i=1}^{n} s_i^A(\hat{\theta}) s_i^A(\hat{\theta})^T - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} s_i^A(\theta^*) s_i^A(\theta^*) \right] \xrightarrow{p} 0$. Consequently, condition (46) implies that $\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} s_i^A(\theta^*) s_i^A(\theta^*) \right] \xrightarrow{p} 0$.

**Proof.** For the first claim, it suffices to show $\frac{n}{\log p}$ and $\sup_{\theta \in S_{1,r}} \left| \frac{1}{n} \sum_{i=1}^{n} s_i^A(\theta) s_i^A(\theta)^T - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} s_i^A(\theta) s_i^A(\theta) \right] \right|_\infty = O(\sqrt{\log n})$. For the second claim, we use the following decomposition

$$M_n \sum_{i=1}^{n} \tilde{H}_i^A = \left[ \mathbb{E} \tilde{H}_i^A \right]^{-1} \sum_{i=1}^{n} \left( \tilde{H}_i^A - H_i^A \right) + \left[ \mathbb{E} \tilde{H}_i^A \right]^{-1} \sum_{i=1}^{n} H_i^A + \left( M_n \left[ \mathbb{E} \tilde{H}_i^A - I^A \right] \right) \left[ \mathbb{E} \tilde{H}_i^A \right]^{-1} \sum_{i=1}^{n} \tilde{H}_i^A.$$

The remaining proofs follow closely the argument for showing Lemma A1 and Lemma A2, and we omit the details here due to the repetitiveness. □

**Lemma A5.** Suppose Assumption 4.2 and the conditions in Lemma 4.1 hold. If $|A| = o_p \left( \sqrt{\frac{n}{(k+1) \log p}} \right)$, we have $\frac{1}{n} \sum_{i=1}^{n} \left[ \theta_j^* \phi'(X_i^\theta) - \theta_j^* \phi'(X_i^\theta^*) \right] = o_p(1)$, where $\theta'$ is some intermediate value $\theta'$ between $\theta^*$ and $\hat{\theta}$.

**Proof.** For some intermediate value $\hat{\theta}$ between $\theta^*$ and $\theta'$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \left[ \theta_j^* \phi'(X_i^\theta) - \theta_j^* \phi'(X_i^\theta^*) \right] = \frac{1}{n} \sum_{i=1}^{n} \theta_j^* \phi''(X_i^\hat{\theta}) X_i^\hat{\theta} \left( \theta' - \theta^* \right).$$

To bound the terms above, we can apply the facts established in previous proofs together with the condition $|A| = o_p \left( \sqrt{\frac{n}{(k+1) \log p}} \right)$. □
B  Additional lemmas

Lemma B1: Let \( \{X_i\}_{i=1}^n \) be independent sub-Exponential random variables with parameter at most \( \sigma \). Then, we have
\[
\mathbb{P}\left[ \left| \sum_{i=1}^n X_i - \mathbb{E}\left( \sum_{i=1}^n X_i \right) \right| \geq n\varepsilon \right] \leq 2 \exp \left( -c_0 n \min \left\{ \frac{\varepsilon^2}{\sigma^2}, \frac{\varepsilon}{\sigma} \right\} \right).
\]

Let \( X \in \mathbb{R}^{n \times p_1} \) be a sub-Gaussian matrix with parameters \((\Sigma_X, \sigma_X^2)\). For any fixed (unit) vector \( \Delta \in \mathbb{R}^{p_1} \), we have
\[
\mathbb{P}\left[ \frac{|X\Delta|^2}{n} - \mathbb{E}\left[ \frac{|X\Delta|^2}{n} \right] \geq \varepsilon \right] \leq 2 \exp \left( -c_1 n \min \left\{ \frac{\varepsilon^2}{\sigma_X^2}, \frac{\varepsilon}{\sigma_X} \right\} \right),
\]
Moreover, if \( Y \in \mathbb{R}^{n \times p_2} \) is a sub-Gaussian matrix with parameters \((\Sigma_Y, \sigma_Y^2)\), then
\[
\mathbb{P}\left[ \frac{Y^T X}{n} - \mathbb{E}(Y_i^T X_i) \geq \varepsilon \right] \leq 6 \exp \left( -c_2 n \min \left\{ \frac{\varepsilon^2}{\sigma_X^2 \sigma_Y^2}, \frac{\varepsilon}{\sigma_X \sigma_Y} \right\} + \log p_1 + \log p_2 \right),
\]
where \( X_i \) and \( Y_i \) are the \( i \)th rows of \( X \) and \( Y \), respectively.

Remark. This lemma is based on Lemma 5.14 and Corollary 5.17 in Vershynin (2012), as well as Lemma 14 in Loh and Wainwright (2012).

Lemma B2: Let \( X \in \mathbb{R}^{n \times p_1} \) be a sub-Gaussian matrix with parameters \((\Sigma_X, \sigma_X^2)\). We have
\[
\frac{|X\Delta|^2}{n} \geq \frac{\alpha}{2} |\Delta|^2 - \alpha \frac{\log p_1 n}{n} |\Delta|^2, \quad \text{for all} \ \Delta \in \mathbb{R}^{p_1}
\]
\[
\frac{|X\Delta|^2}{n} \leq \frac{3\bar{\alpha}}{2} |\Delta|^2 + \alpha' \frac{\log p_1 n}{n} |\Delta|^2, \quad \text{for all} \ \Delta \in \mathbb{R}^{p_1}
\]
with probability at least \( 1 - c_1 \exp(-c_2 n) \), where \( \bar{\alpha}, \bar{\alpha} \), and \( \alpha' \) only depend on \( \Sigma_X \) and \( \sigma_X \).

Remark. This lemma is Lemma 13 in Loh and Wainwright (2012).

Definition (Covering numbers). For a metric space consisting of a set \( \mathcal{X} \) and a metric \( \rho: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \): An \( t \)--covering of \( \mathcal{X} \) with respect to \( \rho \) is a set \( \{\beta^1, ..., \beta^N\} \subset \mathcal{X} \) such that for all \( \beta \in \mathcal{X} \), there exists some \( i \in \{1, ..., N\} \) with \( \rho(\beta, \beta^i) \leq t \). The \( t \)--covering number \( N(t; \mathcal{X}, \rho) \) is the cardinality of the smallest \( t \)--covering.

Lemma B3: For \( q \in (0, 1] \), let
\[
\theta \in \mathcal{B}^d(R_q) := \left\{ \theta' \in \mathbb{R}^d : |\theta'|^q = \sum_{j=1}^d |\theta_j|^q \leq R_q \right\}
\]
and \( N_2(t; \mathcal{B}^d(R_q)) \) be the \( t \)--covering number of the set \( \mathcal{B}^d(R_q) \) in the \( l_2 \)--norm. Then there is a universal constant \( c \) such that
\[
\log N_2(t; \mathcal{B}^d(R_q)) \leq c R_q^{\frac{2}{2-q}} \left( \frac{1}{t} \right)^{\frac{2q}{2-q}} \log d \quad \text{for all} \ t \in (0, R_q^{\frac{1}{2}}).
\]

Remark. These bounds are obtained by inverting known results on (dyadic) entropy numbers (e.g., Schütt, 1984; Guedon and E. Litvak, 2000; Kühn, 2001) of \( l_q \)--“balls” as in the proof for Lemma 2 from Raskutii, Wainwright, and Yu (2011). □
References


