cemmap

centre for microdata methods and practice

Nonlinear factor models for network and panel data

Mingli Chen
Iván Fernández-Val
Martin Weidner

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP38/18

Nonlinear Factor Models for Network and Panel Data*

Mingli Chen[‡] Iván Fernández-Val[§] Martin Weidner[¶]

June 18, 2018

Abstract

Factor structures or interactive effects are convenient devices to incorporate latent variables in panel data models. We consider fixed effect estimation of nonlinear panel single-index models with factor structures in the unobservables, which include logit, probit, ordered probit and Poisson specifications. We establish that fixed effect estimators of model parameters and average partial effects have normal distributions when the two dimensions of the panel grow large, but might suffer of incidental parameter bias. We show how models with factor structures can also be applied to capture important features of network data such as reciprocity, degree heterogeneity, homophily in latent variables and clustering. We illustrate this applicability with an empirical example to the estimation of a gravity equation of international trade between countries using a Poisson model with multiple factors.

Keywords: Panel data, network data, interactive fixed effects, factor models, bias correction, incidental parameter problem, gravity equation

JEL: C13, C23.

*Preliminary versions of this paper were presented at several conferences and seminars. We thank the participants to these presentations, Siyi Luo and Carlo Perroni for helpful comments. Fernández-Val gratefully acknowledges support from the National Science Foundation. Weidner gratefully acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grant ERC-2014-CoG-646917-ROMIA.

‡Department of Economics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK. Email: `m.chen.3@warwick.ac.uk`

§Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215-1403, USA. Email: `ivanf@bu.edu`

¶Department of Economics, University College London, Gower Street, London WC1E 6BT, UK, and CeMMAP. Email: `m.weidner@ucl.ac.uk`

1 Introduction

Factor structures or interactive effects are convenient devices to incorporate latent variables in panel data models. They are commonly used to capture aggregate shocks that might have heterogeneous impacts on the agents in macroeconomic models, and multidimensional individual heterogeneity that might have time varying effects in microeconomic models. More generally, the inclusion of these structures serves to account for dependences along the cross-section and time series dimensions in a parsimonious fashion. While methods for linear factor models are well-established, there are very few studies that develop methods for nonlinear factor models. (We provide a literature review at the end of this section.) Nonlinear models are commonly used when the outcome variable is discrete or has a limited support. In this paper we introduce factor structures in single-index nonlinear specifications such as the logit, probit, ordered probit and Poisson models.

The model that we consider is semiparametric. It includes an outcome, strictly exogenous covariates, and a fixed number of factors and factor loadings. The parametric part is the distribution of the outcome conditional on the covariates, factors and loadings, which is specified up to a finite dimensional parameter. The nonparametric part is the distribution of the factors and loadings conditional on the covariates. In other words, our model is of the "fixed effects" type because we do not impose any restriction on the relationship between the observed covariates and the unobserved factors and loadings. This flexibility allows us to capture features of economic behavior more realistically, but poses important challenges to estimation and inference. The objects of interest are the model parameter and average partial effects (APEs), which are averages of functions of the data, parameter, factors and loadings. The APEs measure the effect of covariates on moments of the outcome conditional on the covariates, factors and loadings. We consider a fixed effects estimation approach that treats the factors and loadings as parameters to be estimated. As it is well-known in the panel data literature, the resulting estimators generally suffer from the incidental parameter problem coming from the high-dimensionality of the estimated parameter (Neyman and Scott, 1948).

We derive asymptotic theory for our estimators of model parameters and APEs under sequences where the two dimensions of the panel pass to infinity with the sample size. Even establishing consistency is complicated in our setting because the dimension of the estimated parameters increases with the sample size. We develop a new proof of consistency that relies on concavity of the log-likelihood function on a single-index that captures the dependence on covariates, parameters, factors and loadings. However, unlike Fernández-Val and Weidner (2016), we need to deal with the complication that our log-likelihood function is not concave in all the estimated parameters because the factors and loadings enter multiplicatively in the index. We also establish that our estimators are normally distributed in large samples, but might have biases of the same order as their standard deviations. For example, we find that the estimator of the model parameter is asymptotically unbiased in the Poisson model, but is biased in logit and probit models. Following the recent panel data literature, we develop analytical and split-sample corrections for the case where the estimator has asymptotic bias. One specific feature of our estimator is that bias depends on the number of factors. In particular, we show that the bias grows proportionally with the number of factors in examples.

We also introduce factor structures as practical tools to model network data. We show how the inclusion of latent factors is useful to incorporate important features of the network such as reciprocity, degree heterogeneity, homophily on latent variables, and clustering (Snijders, 2011; Graham, 2015). We focus on directed networks with unweighted and weighted outcomes. These cover binary response models for network formation where the outcome is an indicator for the existence of a link between the sender and receiver, and count data models for network flow where the outcome is a measure of the volume of flow between the sender and receiver. As we shall discuss, our factor model provides a parsimonious reduced-form specification that captures the important network features mentioned above. The statistical treatment of the network factor model is identical to the panel factor model after noticing that a network is isomorphic to a panel after labeling the senders as individuals and the receivers as time periods.

We illustrate the use of the factor structure in network data with an application to gravity equations of trade between countries. We estimate a Poisson model where the outcome is the volume of trade and the covariates include typical gravity variables such as the distance between the countries or if the country pair belongs to a currency union or a free trade area. The unobserved factors and loadings serve to account for scale and multilateral resistance effects, unobserved partnerships, presence of multinational firms, and differences in natural resources or industrial composition. We find that accounting for these multiple unobserved factors changes the effects of the gravity variables, making all of them to have the expected signs while keeping most of them to be statistically significant.

Literature review: This paper contributes to the econometric panel data and network data literatures. Regarding the panel data literature, our statistical analysis relies on the recent developments in fixed effects methods. We refer to Fernández-Val and Weidner (2017) for a recent review on fixed effects estimation of nonlinear panel models with additive individual and time effects, and to Bai and Wang (2016) for a recent review on fixed effects estimation of linear factor or interactive effects panel models. Since the first draft of this paper appeared in Chen et al. (2014), Boneva and Linton (2017) and Ando and Bai (2016) have considered special cases of nonlinear factor models. Boneva and Linton (2017) analyzed a probit model using the common correlated random effects approach of Pesaran (2006), and Ando and Bai (2016) a logit model using a Bayesian approach with data augmentation. Our analysis is different in the modeling assumptions and estimation method.[1]

In terms of the network literature, our paper is related to the recent work on the application of panel fixed effects methods to network data including Fernández-Val and Weidner (2016), Yan et al. (2016), Cruz-Gonzalez et al. (2017), Dzemski (2017), Graham (2017), and Yan (2018). These papers account for degree heterogeneity by including additive unobserved sender and receiver effects. Additive effects, however, do not capture other network features such as homophily in latent factors and clustering. Graham (2016) considered a binary response model of network formation with all these features plus state dependence, where the network is observed at multiple time periods. Compared to Graham (2016), our method can capture all these features, except for state dependence, applies to binary, ordered and count outcomes, and only requires observing the network at one time period. A stream of the statistic literature has considered nonlinear factor network models using a random effects approach including Hoff et al. (2002), Hoff (2005), Krivitsky et al. (2009), and Handcock et al. (2007). Unlike the fixed effects approach that we adopt, the random effects approach assumes independence between covariates and factors and between covariates and loadings. This assumption is regarded as implausible for most economic applications where the loadings reflect unobserved individual heterogeneity and some of the covariates are individual choice variables. There is also a recent econometric literature on structural models of strategic network formation where the main focus is on identification. We refer to de Paula (2017) for an excellent up-to-date review on this topic. The focus of our paper is on estimation and inference.

Finally, there is an extensive literature in international economics on the estimation of the gravity equation including Harrigan (1994), Eaton and Kortum (2001), Anderson and van Wincoop (2003), Santos Silva and Tenreyro (2006), Helpman et al. (2008), Charbonneau (2012) and Jochmans (2017). We refer to Head and Mayer (2014) for a recent review on this literature. These papers estimate models with additive unobserved sender and receiver country effects to account for scale or multilateral resistence effects. Our innovation to this literature is the inclusion of multiple unobserved factors to account for not only scale effects, but also unobserved partnerships, and homophily induced by differences in natural resource, industrial composition or other country characteristics.

To sum-up, our paper makes the following contributions. First, we derive asymptotic theory for fixed effects estimators of model parameters and APEs in a class of nonlinear single-index

[1] We refer to Boneva and Linton (2017) and Ando and Bai (2016) for more detailed comparisons with our analysis.

factor models that include logit, probit, ordered probit and Poisson models. Second, we provide bias corrections for fixed effect estimators of model parameters and APEs. Third, we bring in the factor structure to model important features of network data such as reciprocity, degree heterogeneity, homophily in latent factors and clustering in a reduced form fashion. Fourth, we apply our methods to the estimation of a gravity equation of trade between countries and confirm the importance of the gravity variables even after conditioning on multiple unobserved latent factors.

Outline: In Section 2, we introduce the model and estimators. Section 3 discusses the statistical issues in the estimation and inference of factor models with a simple example. Section 4 derives asymptotic theory for our estimators. Section 5 describes the results of the empirical application to the gravity equation and a calibrated simulation. The proofs of the main results and other technical details are given in the Appendix.

2 Model and Estimators

2.1 Model

We observe the data $\{(Y_{ij}, X_{ij}) : (i,j) \in \mathcal{D}\}$, where Y_{ij} is a scalar outcome variable and X_{ij} is a d_x-dimensional vector of covariates. The subscripts i and j index individuals and time periods in traditional panels, but they might index different dimensions in other data structures such as network data. In our empirical application, for example, we use country network data where Y_{ij} is the volume of trade between country i and country j, and X_{ij} includes gravity variables such as the distance between country i and country j. Both i and j index countries as exporters and importers respectively. The set $\mathcal{D}$ contains the indexes of the units that are observed. It is a subset of the set of all possible pairs $\mathcal{D}_0 := \{(i,j) : i = 1, \ldots, I; j = 1, \ldots, J\}$, where I and J are the dimensions of the data set. We introduce $\mathcal{D}$ to allow for missing data that are common in panel and network applications. For example, in the trade application $I = J$ and $\mathcal{D} = \mathcal{D}_0 \setminus \{(i,i) : i = 1, \ldots, I\}$ because we do not observe trade of a country with itself. We denote the total number of observations by n, i.e. $n = |\mathcal{D}|$.

We assume that the outcome is generated by

$$Y_{ij} \mid X_{ij}, \beta, \alpha, \gamma \sim f(\cdot \mid z_{ij}), \qquad z_{ij} := X_{ij}'\beta + \pi_{ij}, \qquad \pi_{ij} := \alpha_i' \gamma_j,$$

where f is a known density function with respect to some dominating measure, β is d_x-dimensional parameter vector, and α_i and γ_j are R-vectors of unobserved effects. We collect these effects in the $I \times R$ matrix $\alpha = (\alpha_1, \ldots, \alpha_I)'$, and the $J \times R$ matrix $\gamma = (\gamma_1, \ldots, \gamma_J)'$, which are further stacked in the $R(I+J)$-vector $\phi_n = (\text{vec}(\alpha)', \text{vec}(\gamma)')'$. We make explicit in ϕ_n that the number of unobserved effects changes with the sample size because it will have important effects on estimation and inference. We assume that the dimension of the unobserved effects R is known. The effects α_i and γ_j are unobserved factors and factor loadings. In panel data they represent individual and time effects that in economic applications capture individual heterogeneity and aggregate shocks, respectively. In network data α_i and γ_j represent unobserved characteristics of senders and receivers that affect the network flow. The model is semiparametric because we do not specify the distribution of the unobserved effects nor their relationship with the covariates. This flexibility is important for economic applications where some of the covariates are choice variables with values determined in part by the unobserved effects. The conditional distribution f represents the parametric part of the model.

The model has a single-index specification because the covariates and unobserved effects enter f through the index $z_{ij} = X_{ij}'\beta + \alpha_i'\gamma_j$. The parameter β is a quantity of interest because it measures the effects of the covariates on the distribution of the outcome controlling for the unobserved effects. For example, in network data β can measure homophily in an observable characteristic W if X_{ij} includes $(W_i - W_j)^2$ as one of its components. The unobserved effects have a factor or interactive structure because they enter the index z_{ij} multiplicatively through $\pi_{ij} = \alpha_i'\gamma_j$. The standard additive structure $\alpha_{1i} + \gamma_{1j}$ can be seen as a special case of the factor structure with $R = 2$, $\alpha_i = (\alpha_{1i}, 1)'$, and $\gamma_j = (1, \gamma_{1j})'$. More generally, in panel data applications the factor structure allows one to incorporate multiple aggregate shocks γ_t with heterogeneous effects across agents α_i, or multidimensional individual heterogeneity α_i with time-varying returns γ_t. For example, we can have productivity and monetary shocks with heterogeneous effects across industries, or multiple dimensions of individual ability and skills with time-varying returns in the labor market.

One of the contributions of the paper is to introduce factor structures to network data. In this case the factor structure serves to capture important network features in an unspecified or reduced-form fashion. For example, degree heterogeneity can be captured with the additive structure $\tilde{\alpha}_i + \tilde{\gamma}_t$ mentioned above, and reciprocity by allowing Y_{ij} to be arbitrarily related to Y_{ji} even after conditioning on the covariates and unobserved effects. Another important feature is homophily on latent factors, in addition to the homophily on observed factors captured by X_{ij}. Assume that there is a latent factor ξ_i such that the flow between i and j increases or decreases with the distance between ξ_i and ξ_j as measured by $(\xi_i - \xi_j)^2$. This type of homophily can also be captured by a factor structure with $R = 3$, $\alpha_i = (\xi_i^2, 1, -2\xi_i)'$ and $\gamma_j = (1, \xi_j^2, \xi_j)$. The factor structure can also account for clustering or transitivity of links. Assume that there is a cluster of individuals such that there are more flows within the cluster. This would be captured by a factor structure with $R = 1$, $\alpha_i = \xi_i I_i$ and $\gamma_j = \chi_j I_j$, where ξ_i and χ_j are a positive cluster effects on the sender and receiver, and I_i is an indicator for cluster membership. The factor structure can also account for combinations of these network features. Indeed, one of its advantages is that the researcher has the flexibility of specifying some features and leaving other features unspecified. For example, in the trade application we use a specification that includes additive effects to account explicitly for degree heterogeneity and multiple interactive effects to account for the possibility of having homophily in latent factors and clustering without explicitly modelling any of them.

We consider three running examples throughout the analysis:

Example 1 (Linear model)**.** *Let Y_{ij} be a continuous outcome. We can model the conditional distribution of Y_{ij} using the Gaussian linear model*

$$f(y \mid z_{ij}) = \varphi(z_{ij}/\sigma)/\sigma, \quad y \in \mathbb{R},$$

where φ is the density function of the standard normal and σ is a positive scale parameter.

Example 2 (Binary response model)**.** *Let Y_{ij} be a binary outcome and F be a cumulative distribution function of the standard normal or logistic distribution. We can model the conditional distribution of Y_{ij} using the probit or logit model*

$$f(y \mid z_{ij}) = F(z_{ij})^y[1 - F(z_{ij})]^{1-y}, \quad y \in \{0, 1\}.$$

Example 3 (Count response model)**.** *Let Y_{ij} be a count or non-negative integer-valued outcome, and $\psi(\cdot; \lambda)$ be the probability mass function of a Poisson random variable with parameter $\lambda > 0$. We can model the conditional distribution of Y_{ij} using the Poisson model*

$$f(y \mid z_{ij}) = \psi(y; \exp[z_{ij}]), \quad y \in \{0, 1, 2, ...\}.$$

2.2 Average Partial Effects

In addition to the model parameter β, we might be interested in average partial effects (APEs). These effects are averages of the data, parameters and unobserved effects. They measure the effect of the covariates on moments of the distribution of the outcome conditional on the covariates and unobserved effects. The leading case is the conditional expectation,

$$\mathbb{E}[Y_{ij} \mid X_{ij}, \alpha_i, \gamma_j, \beta] = \int y f(y \mid X_{ij}'\beta + \pi_{ij})dy,$$

where the partial effects are differences or derivatives of this expression with respect to the components of X_{ij}. We denote generically the partial effects by $\Delta(Y_{ij}, X_{ij}, \beta, \alpha_i'\gamma_j) = \Delta_{ij}(\beta, \alpha_i'\gamma_j)$, where the restriction that they depend on α_i and γ_j through π_{ij} is natural given the model for the conditional density of Y_{ij}. We allow the partial effect to depend on Y_{ij} to cover scale and other parameters not included in the single-index. The APE is

$$\delta = \mathbb{E}\left[\frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \Delta_{ij}(\beta, \alpha_i'\gamma_j)\right]. \tag{2.1}$$

Example 1 (Linear model). *The variance σ^2 in the linear model can be expressed as an APE with*

$$\Delta_{ij}(\beta, \alpha_i'\gamma_j) = (Y_{ij} - X_{ij}'\beta - \alpha_i'\gamma_j)^2. \tag{2.2}$$

Example 2 (Binary response model). *If $X_{ij,k}$, the kth element of X_{ij}, is binary, its partial effect on the conditional probability of Y_{ij} is*

$$\Delta_{ij}(\beta, \alpha_i'\gamma_j) = F(\beta_k + X_{ij,-k}'\beta_{-k} + \alpha_i'\gamma_j) - F(X_{ij,-k}'\beta_{-k} + \alpha_i'\gamma_j), \tag{2.3}$$

where β_k is the kth element of β, and $X_{ij,-k}$ and β_{-k} include all elements of X_{ij} and β except for the kth element. If $X_{ij,k}$ is continuous and F is differentiable, the partial effect of $X_{ij,k}$ on the conditional probability of Y_{ij} is

$$\Delta_{ij}(\beta, \alpha_i'\gamma_j) = \beta_k \partial F(X_{ij}'\beta + \alpha_i'\gamma_j), \quad \partial F(u) := \partial F(u)/\partial u. \tag{2.4}$$

Example 3 (Count response model). *If $X_{ij,k}$, the kth element of X_{ij}, is binary, its partial effect on the conditional probability of Y_{ij} in the Poisson model is*

$$\Delta_{ij}(\beta, \alpha_i'\gamma_j) = \exp(\beta_k + X_{ij,-k}'\beta_{-k} + \alpha_i'\gamma_j) - \exp(X_{ij,-k}'\beta_{-k} + \alpha_i'\gamma_j), \tag{2.5}$$

where β_k is the kth element of β, and $X_{ij,-k}$ and β_{-k} include all elements of X_{ij} and β except for the kth element. If $X_{ij,k}$ is continuous, the partial effect of $X_{ij,k}$ on the conditional expectation of Y_{ij} is

$$\Delta_{ij}(\beta, \alpha_i'\gamma_j) = \beta_k \exp(X_{ij}'\beta + \alpha_i'\gamma_j). \tag{2.6}$$

2.3 Fixed effects estimator

We adopt a fixed effect approach and treat the unobserved effects ϕ_n as a vector of nuisance parameters to be estimated. Let

$$L(\theta, \phi_n) := \sum_{(i,j)\in\mathcal{D}} \log f(Y_{ij} \mid X_{ij}'\beta + \pi_{ij}),$$

be the conditional log-likelihood function of the data constructed from the parametric part of the model. The fixed-effect estimator is

$$(\widehat{\beta}, \widehat{\phi}_n) \in \underset{(\beta,\phi_n)\in\mathbb{R}^{d_x+R(I+J)}}{\operatorname{argmax}} L(\beta, \phi_n). \tag{2.7}$$

This problem has unique solution with probability one for β under the assumption that $z \mapsto \log f(\cdot \mid z)$ is concave. This assumption holds for all the cases that we consider including logit, probit, ordered probit and Poisson models. The solution for ϕ_n is only unique up to the standard normalizations for linear factor models.[2] Obtaining the solution to (2.7) can be computationally challenging because the objective function is not concave in the parameter ϕ_n and the high-dimensionality of the parameter space. In the numerical examples of Section 5, we use the iterative method of Chen (2014) based on the EM algorithm. This method performs well in simulations.

Plugging the estimator of (β, ϕ_n) in (2.1) yields the estimator of the APE,

$$\widehat{\delta} = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Delta_{ij}(\widehat{\beta}, \widehat{\alpha}_i'\widehat{\gamma}_j). \tag{2.8}$$

This estimator is invariant to the normalization used to pin down the value of $\widehat{\phi}_n$ in (2.7). In Section 4, we show that $\widehat{\beta}$ and $\widehat{\delta}$ are consistent and normally distributed in large samples, but might have incidental parameter bias because the dimension of the nuisance parameter ϕ_n grows with the sample size (Neyman and Scott, 1948).

3 A Simple Motivating Example

We illustrate the statistical issues that arise in the estimation of factor models with a simple example. The analysis in this section is mainly heuristic leaving technical details such as the derivation of the orders of some remainder terms in the asymptotic expansions for Section 4.

Consider a version of Example 1 without covariates where $Y_{ij} \mid \phi_n \sim \mathcal{N}(\alpha_i'\gamma_j, \sigma^2)$. Assume that the observations Y_{ij} are independent over i and j, and that there is no missing data, i.e. $\mathcal{D} = \mathcal{D}_0$. The quantity of interest is the scale parameter σ^2, which can be treated as a APE. This is a linear factor model where $\widehat{\phi}_n$ can be obtained using the principal component algorithm of Bai (2009). Then, the plug-in estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \widehat{\alpha}_i'\widehat{\gamma}_j\right)^2. \tag{3.1}$$

[2]The solution for ϕ_n is not uniquely determined because the log-likelihood function is invariant to transformations $\alpha \mapsto \alpha A'$ and $\gamma \mapsto \gamma A^{-1}$ for any non-singular $R \times R$ matrix A.

To analyze the properties of $\widehat{\sigma}^2$, it is useful to consider an asymptotic expansion of $\widehat{\alpha}_i'\widehat{\gamma}_j$ around $\alpha_i'\gamma_j$ as $I, J \to \infty$. This yields

$$\widehat{\alpha}_i'\widehat{\gamma}_j = \alpha_i'\gamma_j + (\widehat{\alpha}_i - \alpha_i)'\gamma_j + \alpha_i'(\widehat{\gamma}_j - \gamma_j) + (\widehat{\alpha}_i - \alpha_i)'(\widehat{\gamma}_j - \gamma_j)$$
$$\approx \alpha_i'\gamma_j + (\widehat{\alpha}_i - \alpha_i)'\gamma_j + \alpha_i'(\widehat{\gamma}_j - \gamma_j),$$

where $\approx$ means equal up to terms of lower order. Plugging this expansion in (3.1) shows that $\widehat{\sigma}^2$ behaves asymptotically as a sample variance with $R(I+J)$ estimated fixed effects corresponding to the $\widehat{\alpha}_i$'s and $\widehat{\gamma}_t$'s. Then, standard degrees of freedom calculations give

$$\mathbb{E}[\widehat{\sigma}^2] \approx \frac{(I-R)(J-R)}{IJ}\sigma^2 \approx \sigma^2 - \frac{R(I+J)}{IJ}\sigma^2, \tag{3.2}$$

which shows that $\widehat{\sigma}^2$ has an incidental parameter bias that grows proportionally to the number of factors R. The order of the bias corresponds to the number of estimated parameters, $R(I+J)$, divided by the number of observations, IJ, as predicted by the general formula in Fernández-Val and Weidner (2017) for fixed-effects estimators. We show in numerical examples that this expression produces a very accurate approximation to the bias even for small sample sizes.

We carry out 50,000 simulations with $\sigma^2 = 1$, and α_i and γ_j drawn independently from multivariate normal distributions with mean zero and covariate function $\mathbb{I}_R$, the identity matrix of order R. Table 1 compares the bias of $\widehat{\sigma}^2$ with the asymptotic approximation (3.2) in datasets with $I, J \in \{10, 25, 50\}$, and $R \in \{1, 2, 3\}$. We only report the results for $J \leq I$ since all the expressions are symmetric in I and J. Comparing the two rows in each panel of the table, we find that the asymptotic bias provides a very accurate approximation to the finite-sample bias of the estimator for all the sample sizes and numbers of factors.

The bias of $\widehat{\sigma}^2$ can be removed using analytical and split-sample methods. Thus, an analytical bias corrected estimator can be formed as

$$\widetilde{\sigma}^2_{ABC} = \frac{IJ}{(I-R)(J-R)}\widehat{\sigma}^2.$$

A split-sample bias corrected estimator can be formed as

$$\widetilde{\sigma}^2_{SBC} = 3\widehat{\sigma}^2 - \bar{\sigma}^2_{I,J/2} - \bar{\sigma}^2_{I/2,J},$$

where $\bar{\sigma}^2_{I,J/2}$ is the average of the estimators in the haft-panels $\{(i,j) : i = 1, \dots, I; j = 1, \dots, \lceil J/2 \rceil\}$ and $\{(i,j) : i = 1, \dots, I; j = \lfloor J/2+1 \rfloor, \dots, J\}$, and $\bar{\sigma}^2_{I/2,J}$ is the average of the estimators in the haft-panels $\{(i,j) : i = 1, \dots, \lceil I/2 \rceil; j = 1, \dots, J\}$ and $\{(i,j) : i = \lfloor I/2+1 \rfloor, \dots, I; j = 1, \dots, J\}$, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceil and floor functions. As in nonlinear panel data, we expect these corrections to remove most of the bias of the estimator without increasing dispersion. Moreover,

Table 1: Asymptotic and Exact Bias of $\widehat{\sigma}^2$

	$I=10$	$I=25$		$I=50$		
Bias	$J=10$	$J=10$	$J=25$	$J=10$	$J=25$	$J=50$
			$R=1$			
Asymptotic	-.19	-.14	-.08	-.12	-.06	-.04
Exact	-.20	-.14	-.08	-.12	-.06	-.04
			$R=2$			
Asymptotic	-.36	-.26	-.15	-.23	-.12	-.08
Exact	-.39	-.27	-.16	-.24	-.12	-.08
			$R=3$			
Asymptotic	-.51	-.38	-.23	-.34	-.17	-.12
Exact	-.55	-.40	-.23	-.35	-.18	-.12

Notes: Results obtained by 50,000 simulations

Design: $Y_{ij} \mid \phi_n \sim \mathcal{N}(\alpha_i'\gamma_t, \sigma^2)$, $\alpha_i \sim N(0, \mathbb{I}_R)$, $\gamma_j \sim N(0, \mathbb{I}_R)$, $\sigma^2 = 1$

Table 2: Bias, SD, RMSE and Coverage Probabilities

	Bias	SD	RMSE	Cover	Bias	SD	RMSE	Cover
		$I=10, J=10$				$I=25, J=10$		
$\widehat{\sigma}^2$	-0.55	0.09	0.56	0.00	-0.40	0.07	0.41	0.00
$\widetilde{\sigma}^2_{ABC}$	-0.08	0.19	0.20	0.75	-0.02	0.11	0.11	0.85
$\widetilde{\sigma}^2_{SBC}$	-0.09	0.20	0.22	0.71	-0.03	0.12	0.13	0.81
		$I=25, J=25$				$I=50, J=10$		
$\widehat{\sigma}^2$	-0.23	0.05	0.24	0.01	-0.35	0.05	0.35	0.00
$\widetilde{\sigma}^2_{ABC}$	-0.01	0.06	0.06	0.91	-0.01	0.08	0.08	0.88
$\widetilde{\sigma}^2_{SBC}$	-0.02	0.07	0.07	0.85	-0.01	0.08	0.08	0.85
		$I=50, J=25$				$I=50, J=50$		
$\widehat{\sigma}^2$	-0.18	0.04	0.18	0.00	-0.12	0.03	0.12	0.01
$\widetilde{\sigma}^2_{ABC}$	-0.00	0.04	0.04	0.92	-0.00	0.03	0.03	0.93
$\widetilde{\sigma}^2_{SBC}$	-0.01	0.05	0.05	0.88	-0.00	0.03	0.03	0.92

Notes: 50,000 simulations. Nominal level is 0.95

Design: $Y_{ij} \mid \phi_n \sim \mathcal{N}(\alpha_i'\gamma_t, \sigma^2)$, $\sigma^2 = 1$, $\alpha_i \sim N(0, \mathbb{I}_R)$, $\gamma_j \sim N(0, \mathbb{I}_R)$, $R = 3$

constructing confidence intervals around the corrected estimators should help bring coverage probabilities close to their nominal levels. We confirm these predictions in a numerical simulation.

Table 2 reports the bias, standard deviation and rmse of the uncorrected and bias corrected estimators, together with coverage probabilities of 95% confidence interval constructed around them. The results are based on 50,000 simulations of datasets generated as in Table 1 with $I, J \in \{10, 25, 50\}$, and $R = 3$. The confidence intervals around the estimator $\widetilde{\sigma}^2 \in \{\widehat{\sigma}^2, \widetilde{\sigma}^2_{ABC} \widetilde{\sigma}^2_{SBC}\}$ are constructed as $\widetilde{\sigma}^2(1 \pm 1.96\sqrt{2/(IJ)})$, where we use that the asymptotic variance of all the estimators is $2\sigma^4/(IJ)$. We find that the corrections offer huge improvements in terms of bias reduction and coverage of the confidence intervals. The corrections increase the dispersion for small sample sizes, but always reduce the rmse. In this case the analytical correction slightly overperforms the split-sample correction.

4 Asymptotic Theory

We derive the asymptotic distribution of the estimators of the model parameter and APEs under sequences where I and J grow with the sample size at the same rate. We focus on these sequences because they are the only ones that deliver a non-degenerate limit distribution. Moreover, they are very natural choices for network data where $I = J$. Throughout this section, all the stochastic statements are conditional on the realization of the unobserved effects ϕ_n and should therefore be qualified with almost surely. We shall omit this qualifier to lighten the notation.

4.1 Model parameter

We consider single-index models with strictly exogenous covariates and unobserved effects that enter the density of the outcome through $z_{ij} = X_{ij}'\beta + \pi_{ij}$, where $\pi_{ij} = \alpha_i'\gamma_j$. These models cover the linear, probit and Poisson specifications of Examples 1–3. We focus on strictly exogenous covariates because for some data structures of interest such as network data there is no natural ordering of the observations. The results can be extended to predetermined covariates when one of the dimensions is time, see the earlier version of the paper (Chen et al., 2014). Let

$$\ell_{ij}(z_{ij}) := \log f(Y_{ij} \mid X_{ij}, \beta, \alpha_i, \gamma_j) \tag{4.1}$$

be the conditional log-likelihood coming from the parametric part of the model. We denote the derivatives of $z \mapsto \ell_{ij}(z)$ by $\partial_{z^q}\ell_{ij}(z) := \partial^q \ell_{ij}(z)/\partial z^q$, $q = 1, 2, \ldots$. Let β^0, α_i^0, γ_j^0, and $\pi_{ij}^0 = \alpha_i^{0\prime}\gamma_j^0$ denote the values of β, α_i, γ_j, and π_{ij} that generated the data. We drop the argument z_{ij} when the derivatives are evaluated at the true value of the index $z_{ij}^0 := X_{ij}'\beta^0 + \pi_{ij}^0$, i.e., $\partial_{z^q}\ell_{ij} := \partial_{z^q}\ell_{ij}(z_{ij}^0)$. Let $\boldsymbol{X} = \{X_{ij} : (i, j) \in \mathcal{D}\}$.

We make the following assumptions:

Assumption 1 (Nonlinear Factor Model)**.** *Let $\varepsilon > 0$ and let $\mathcal{B}^0_\varepsilon$ be a bounded subset of $\mathbb{R}$ that contains an ε-neighborhood of z^0_{ij} for all i, j, I, J.*

(i) Model: conditional on $\boldsymbol{X}$, Y_{ij} is distributed as

$$Y_{ij} \sim \exp[\ell_{ij}(X'_{ij}\beta^0 + \pi^0_{ij})],$$

and either (a) Y_{ij} is independent across $(i, j) \in \mathcal{D}$ or (b) (Y_{ij}, Y_{ji}) is independent across observations $(i, j) \in \mathcal{D}$ with $i \leq j$. The number of factors R is known.

(ii) Asymptotics: we consider limits of sequences where $I_n/J_n \to \kappa^2$, $0 < \kappa < \infty$, as $n = |\mathcal{D}| \to \infty$. We shall drop the indexing by n from I_n and J_n.

(iii) Smoothness and moments: $z \mapsto \ell_{ij}(z)$ is four times continuously differentiable over $\mathcal{B}^0_\varepsilon$ a.s. and $\max_{i,j} \mathbb{E}[|\partial_{z^q}\ell_{ij}(z^0_{ij})|^{8+\nu}]$, $q \leq 4$, are uniformly bounded over I, J for some $\nu > 0$. In addition, X_{ij} is bounded uniformly over i, j, I, J.

(iv) Concavity: for all I, J, the function $z \mapsto \ell_{ij}(z)$ is strictly concave over $z \in \mathbb{R}$ a.s. Furthermore, there exist positive constants $b_{\min}$ and $b_{\max}$ such that for all $z \in \mathcal{B}^0_\varepsilon$, $b_{\min} \leq -\partial_{z^2}\ell_{ij}(z) \leq b_{\max}$ a.s. uniformly over i, j, I, J.

(v) Strong factors: $I^{-1}\sum_{i=1}^I \alpha^0_i \alpha^{0\prime}_i \to_P \Sigma_1 > 0$, and $J^{-1}\sum_j \gamma^0_j \gamma^{0\prime}_j \to_P \Sigma_2 > 0$.

(vi) Generalized non-collinearity: for any matrix A, define the coprojection matrix as $\mathcal{M}_A := \mathbb{I} - A(A'A)^\dagger A'$, where $\mathbb{I}$ denotes the identity matrix of appropriate dimensions and the superscript † denotes the Moore-Penrose generalized inverse. Let $\alpha^0 := (\alpha^0_1, \ldots, \alpha^0_I)'$ and $\mathbb{X}_k$ be a $I \times J$ matrix with elements $X_{ij,k}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. The $d_x \times d_x$ matrix $D(\gamma)$ with elements

$$D_{k_1 k_2}(\gamma) = (IJ)^{-1}\mathrm{Tr}(\mathcal{M}_{\alpha^0}\mathbb{X}_{k_1}\mathcal{M}_\gamma \mathbb{X}'_{k_2}), \quad k_1, k_2 \in \{1, ..., d_x\},$$

satisfies $D(\gamma) > c > 0$ for all $\gamma \in \mathbb{R}^{J \times R}$, wpa1.

(vii) Missing data: there is a finite number of missing observations for every i and j, that is, $\max_i(J - |\{(i', j') \in \mathcal{D} : i' = i\}|) \leq C$ and $\max_j(I - |\{(i', j') \in \mathcal{D} : j' = j\}|) \leq C$ for some constant $C < \infty$ that is independent of the sample size.

The two cases considered in Assumption 1(i) are designed for different data structures. Case (b) is more suitable for network data because it allows for reciprocity between the observations (i, j) and (j, i), whereas case (a) is more suitable for panel data where there is no special relationship between these observations. Assumption 1(i) also imposes that the number of factors is known. In practice, we recommend checking the sensitivity to this number by reporting the maximum value of the average log-likelihood and the parameter estimates for multiple values of R. We provide an example in the empirical application of Section 5.[3] Assumption 1$(i)-(iii)$ are similar to Fernández-Val and Weidner (2016), so we do not discuss them further here. The concavity condition in Assumption 1(iv) holds for the logit, probit, ordered probit and Poisson models. The strong factor and generalized noncollinearity conditions in Assumption 1$(v)-(vi)$ were previously imposed in Bai (2009) and Moon and Weidner (2015, 2017) for linear models with interactive effects. Generalized noncollinearity rules out covariates that do not display variation in the two dimensions of the dataset. Boneva and Linton (2017) and Ando and Bai (2016) impose very similar conditions to Assumption 1, so we refer to these papers for further discussion.

We introduce some notation that is convenient to simplify the expressions in the asymptotic distribution. Let Ξ_{ij} be a d_x-dimensional vector defined by the following population weighted least squares projection for each component of $\mathbb{E}(\partial_{z^2}\ell_{ij}X_{ij})$,

$$\Xi_{it,k} = \alpha_{i,k}^{*\prime}\gamma_j^0 + \alpha_i^{0\prime}\gamma_{t,k}^*, \quad (\alpha_k^*, \gamma_k^*) = \operatorname*{argmin}_{\alpha_{i,k},\gamma_{t,k}} \sum_{i,j} \mathbb{E}(-\partial_{z^2}\ell_{ij}) \left(\frac{\mathbb{E}(\partial_{z^2}\ell_{ij}X_{ij,k})}{\mathbb{E}(\partial_{z^2}\ell_{ij})} - \alpha_{i,k}'\gamma_j^0 - \alpha_i^{0\prime}\gamma_{t,k} \right)^2.$$

Also define the residual of the projection

$$\tilde{X}_{ij} := X_{ij} - \Xi_{ij}.$$

Finally, let $\overline{\mathbb{E}} := \operatorname{plim}_{I,J\to\infty}$, $\mathcal{D}_i := \{j : (i,j) \in \mathcal{D}\}$ and $\mathcal{D}_j := \{i : (i,j) \in \mathcal{D}\}$.

The following theorem establishes the asymptotic distribution of $\widehat{\beta}$ defined in (2.7).

Theorem 1 (Asymptotic distribution of $\widehat{\beta}$)**.** *Suppose that Assumption 1 holds, that the following limits exist*

$$\overline{B}_\infty = -\overline{\mathbb{E}} \left\{ \frac{1}{I} \sum_{(i,j)\in\mathcal{D}} \gamma_j^{0\prime} \left[\sum_{h\in\mathcal{D}_i} \gamma_h^0 \gamma_h^{0\prime}\, \mathbb{E}\left(\partial_{z^2}\ell_{ih}\right) \right]^{-1} \gamma_j^0\, \mathbb{E}\left(\partial_z \ell_{ij} \partial_{z^2}\ell_{ij} \tilde{X}_{ij} + \frac{1}{2}\partial_{z^3}\ell_{ij}\tilde{X}_{ij} \right) \right\},$$

$$\overline{D}_\infty = -\overline{\mathbb{E}} \left\{ \frac{1}{J} \sum_{(i,j)\in\mathcal{D}} \alpha_i^{0\prime} \left[\sum_{h\in\mathcal{D}_j} \alpha_h^0 \alpha_h^{0\prime}\, \mathbb{E}\left(\partial_{z^2}\ell_{hj}\right) \right]^{-1} \alpha_i^0\, \mathbb{E}\left(\partial_z \ell_{ij} \partial_{z^2}\ell_{ij} \tilde{X}_{ij} + \frac{1}{2}\partial_{z^3}\ell_{ij}\tilde{X}_{ij} \right) \right\},$$

$$\overline{W}_\infty = -\overline{\mathbb{E}} \left[\frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left(\partial_{z^2}\ell_{ij} \tilde{X}_{ij} \tilde{X}_{ij}' \right) \right],$$

$$\overline{\Sigma}_\infty = \overline{\mathbb{E}} \left[\frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left\{ \left(\partial_z\ell_{ij}\tilde{X}_{ij} + \partial_z\ell_{ji}\tilde{X}_{ji} \right) \partial_z\ell_{ij}\tilde{X}_{ij}' \right\} \right],$$

[3] It would be of interest to develop a method to select R similar to the criteria of Bai and Ng (2002) or Ando and Bai (2016). We leave this extension to future research.

and that $\overline{W}_\infty > 0$. Then,

$$\sqrt{n}\left(\widehat{\beta} - \beta^0 - \frac{I}{n}\,\overline{W}_\infty^{-1}\overline{B}_\infty - \frac{J}{n}\,\overline{W}_\infty^{-1}\overline{D}_\infty\right) \to_d \mathcal{N}(0,\ \overline{W}_\infty^{-1}\overline{\Sigma}_\infty\overline{W}_\infty^{-1}).$$

Remark 1 (Panel Data). *In case (a) of Assumption 1(i), the asymptotic variance of $\widehat{\beta}$ simplifies to*

$$\overline{W}_\infty^{-1}\overline{\Sigma}_\infty\overline{W}_\infty^{-1} = -\overline{W}_\infty^{-1},$$

by the fact that the scores $\partial_z \ell_{ij}\tilde{X}_{ij}$ and $\partial_z \ell_{ji}\tilde{X}_{ji}$ are uncorrelated and the information equality.

Theorem 1 shows that $\widehat{\beta}$ is consistent and normally distributed, but can have bias of the same order as its standard deviation. The scaling factor in the expressions for $\overline{B}_\infty$ and $\overline{D}_\infty$ are such that those expressions are of order one, for example, we can express $\overline{B}_\infty$ equivalently as

$$-\overline{\mathbb{E}}\left\{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{|\mathcal{D}_i|}\sum_{j\in\mathcal{D}_i}\gamma_j^{0\,\prime}\left[\frac{1}{|\mathcal{D}_i|}\sum_{h\in\mathcal{D}_i}\gamma_h^0\gamma_h^{0\prime}\,\mathbb{E}\left(\partial_{z^2}\ell_{ih}\right)\right]^{-1}\gamma_j^0\,\mathbb{E}\left(\partial_z\ell_{ij}\partial_{z^2}\ell_{ij}\tilde{X}_{ij} + \frac{1}{2}\partial_{z^3}\ell_{ij}\tilde{X}_{ij}\right)\right\},$$

where all sums explicitly appear as part of a sample average. We verify the presence of bias in our running examples.

Example 1 (Linear model). *In this case*

$$\ell_{ij}(z) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_{ij} - z_{ij})^2}{2\sigma^2},$$

so that $\partial_z\ell_{ij} = (Y_{ij} - z_{ij}^0)/\sigma^2$, $\partial_{z^2}\ell_{ij} = -1/\sigma^2$, and $\partial_{z^3}\ell_{ij} = 0$. Substituting these values in the expressions of the bias of Theorem 1 yields $\overline{B}_\infty = \overline{D}_\infty = 0$, which agrees with the result in Bai (2009) of no asymptotic bias in homoskedastic linear models with interactive effects and strictly exogenous covariates.

Example 2 (Binary response model). *In this case*

$$\ell_{ij}(z) = Y_{ij}\log F(z) + (1 - Y_{ij})\log[1 - F(z)],$$

so that $\partial_z\ell_{ij} = H_{ij}(Y_{ij} - F_{ij})$, $\partial_{z^2}\ell_{ij} = -H_{ij}\partial F_{ij} + \partial H_{ij}(Y_{ij} - F_{ij})$, and $\partial_{z^3}\ell_{ij} = -H_{ij}\partial^2 F_{ij} - 2\partial H_{ij}\partial F_{ij} + \partial^2 H_{ij}(Y_{ij} - F_{ij})$, where $H_{ij} = \partial F_{ij}/[F_{ij}(1-F_{ij})]$, and $\partial^j G_{ij} := \partial^j G(Z)|_{Z=z_{ij}^0}$ for any function G and $j = 0, 1, 2$. Substituting these values in the expressions of the bias of Theorem 1 for the probit model yields

$$\overline{B}_\infty = \overline{\mathbb{E}}\left\{\frac{1}{2\,I}\sum_{(i,j)\in\mathcal{D}}\gamma_j^{0\,\prime}\left[\sum_{h\in\mathcal{D}_i}\gamma_h^0\gamma_h^{0\prime}\,\mathbb{E}\left(\partial_{z^2}\ell_{ih}\right)\right]^{-1}\gamma_j^0\,\mathbb{E}\left(\partial_{z^2}\ell_{ij}\tilde{X}_{ij}\tilde{X}_{ij}'\right)\right\}\beta^0,$$

$$\overline{D}_\infty = \overline{\mathbb{E}}\left\{\frac{1}{2\,J}\sum_{(i,j)\in\mathcal{D}}\alpha_i^{0\,\prime}\left[\sum_{h\in\mathcal{D}_j}\alpha_h^0\alpha_h^{0\prime}\,\mathbb{E}\left(\partial_{z^2}\ell_{hj}\right)\right]^{-1}\alpha_i^0\,\mathbb{E}\left(\partial_{z^2}\ell_{ij}\tilde{X}_{ij}\tilde{X}_{ij}'\right)\right\}\beta^0.$$

The asymptotic bias is therefore a positive definite matrix weighted average of the true parameter value as in the case of the probit model with additive individual and time effects in Fernández-Val and Weidner (2016). The bias grows linearly with the number of factors because

$$\sum_{j\in\mathcal{D}_i}\gamma_j^{0\,\prime}\left[\sum_{h\in\mathcal{D}_i}\gamma_h^0\gamma_h^{0\prime}\right]^{-1}\gamma_j^0=\sum_{i\in\mathcal{D}_j}\alpha_i^{0\,\prime}\left[\sum_{h\in\mathcal{D}_j}\alpha_h^0\alpha_h^{0\prime}\right]^{-1}\alpha_i^0=R, \tag{4.2}$$

and $\mathbb{E}\left(\partial_{z^2}\ell_{ij}\right)$ *and* $\mathbb{E}\left(\partial_{z^2}\ell_{ij}\tilde{X}_{ij}\tilde{X}_{ij}'\right)$ *are bounded uniformly in* i,j.

Example 3 (Count response model). *In this case*

$$\ell_{ij}(z)=zY_{ij}-\exp(z)-\log Y_{ij}!,$$

so that $\partial_z\ell_{ij}=Y_{ij}-\lambda_{ij}$ *and* $\partial_{z^2}\ell_{ij}=\partial_{z^3}\ell_{ij}=-\lambda_{ij}$*, where* $\lambda_{ij}=\exp(z_{ij}^0)$*. Substituting these values in the expressions of the bias of Theorem 1 yields*

$$\overline{B}_\infty=\overline{D}_\infty=0,$$

which generalizes the result in Fernández-Val and Weidner (2016) of no asymptotic bias in the Poisson model with strictly exogenous covariates and additive individual and time effects to the Poisson model with strictly exogenous covariates and multiple factors.

4.2 Average Partial Effects

We use additional assumptions to derive the asymptotic distribution of the estimator of the APEs. They involve smoothness conditions on the partial effect function $(\beta,\pi)\mapsto\Delta_{ij}(\beta,\pi)$ needed to obtain the limit distribution of $\widehat{\delta}$ from the limit distribution of $(\widehat{\beta},\widehat{\phi}_n)$ via delta method. For a vector of nonnegative integer numbers $v=(v_1,\ldots,v_{d_x})$, let $\partial_{\beta^v}:=\partial^{|v|}/\partial\beta_1^{v_1}\cdots\partial\beta_{d_x}^{v_{d_x}}$ and $|v|=v_1+\ldots+v_{d_x}$.

Assumption 2 (Partial effects)**.** *Let* $\epsilon>0$*, and let* $\mathcal{B}_\varepsilon^0$ *be a subset of* $\mathbb{R}^{d_x+1}$ *that contains an* ε*-neighborhood of* (β^0,π_{ij}^0) *for all* i,j,I,J*..*

(i) Model: for all i,j,I,J*, the partial effects depend on* α_i *and* γ_j *through* $\pi_{ij}=\alpha_i'\gamma_j$*:*

$$\Delta(Y_{ij},X_{ij},\beta,\alpha_i,\gamma_j)=\Delta_{ij}(\beta,\pi_{ij}),$$

where $(\beta,\pi)\mapsto\Delta_{ij}(\beta,\pi)$ *is a known real-valued function. The realizations of the partial effects are denoted by* $\Delta_{ij}:=\Delta_{ij}(\beta^0,\pi_{ij}^0)$*.*

(ii) Smoothness and moments: The function $(\beta,\pi)\mapsto\Delta_{ij}(\beta,\pi)$ *is four times continuously differentiable over* $\mathcal{B}_\varepsilon^0$ *a.s., and* $\max_{i,j}\mathbb{E}[|\partial_{\beta^v\pi^q}\ell_{ij}(\beta^0,z_{ij}^0)|^{8+\nu}]$*,* $|v|+q\le 4$*, are uniformly bounded over* I,J *for some* $\nu>0$*.*

It is convenient again to introduce some notation to simplify the expressions of the asymptotic distribution. Let Ψ_{ij} be weighted least squares population projection

$$\Psi_{ij} = \alpha_i^{*\prime}\gamma_j^0 + \alpha_i^{0\prime}\gamma_t^*, \qquad (\alpha^*, \gamma^*) = \underset{\alpha_i,\gamma_t}{\operatorname{argmin}} \sum_{i,j} \mathbb{E}(-\partial_{z^2}\ell_{ij}) \left(\frac{\mathbb{E}(\partial_\pi \Delta_{ij})}{\mathbb{E}(\partial_{z^2}\ell_{ij})} - \alpha_i'\gamma_j^0 - \alpha_i^{0\prime}\gamma_t \right)^2.$$

We denote the partial derivatives of $(\beta,\pi) \mapsto \Delta_{ij}(\beta,\pi)$ by $\partial_\beta \Delta_{ij}(\beta,\pi) := \partial\Delta_{ij}(\beta,\pi)/\partial\beta$, $\partial_{\beta\beta'}\Delta_{ij}(\beta,\pi) := \partial^2\Delta_{ij}(\beta,\pi)/(\partial\beta\partial\beta')$, $\partial_{\pi^q}\Delta_{ij}(\beta,\pi) := \partial^q\Delta_{ij}(\beta,\pi)/\partial\pi^q$, $q = 1,2,3,\ldots$. We drop the arguments β and π when the derivatives are evaluated at the true values β^0 and π_{ij}^0, e.g. $\partial_{\pi^q}\Delta_{ij} := \partial_{\pi^q}\Delta_{ij}(\beta^0,\pi_{ij}^0)$. We also define $D_\pi\Delta_{ij} := \partial_\pi\Delta_{ij} - \partial_{z^2}\ell_{ij}\Psi_{ij}$ and $D_{\pi^2}\Delta_{ij} := \partial_{\pi^2}\Delta_{ij} - \partial_{z^3}\ell_{ij}\Psi_{ij}$.

We are now ready to present the asymptotic distribution of $\widehat{\delta}$ defined in (2.8).

Theorem 2 (Asymptotic distribution of $\widehat{\delta}$)**.** *Suppose that the assumptions of Theorem 1 and Assumption 2 hold, and that the following limits exist:*

$$\overline{(D_\beta\Delta)}_\infty = \overline{\mathbb{E}}\left[\frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \mathbb{E}(\partial_\beta\Delta_{ij} - \Xi_{ij}\partial_\pi\Delta_{ij})\right]',$$

$$\overline{B}_\infty^\delta = -\overline{\mathbb{E}}\left\{\frac{1}{I}\sum_{(i,j)\in\mathcal{D}} \gamma_j^{0\prime}\left[\sum_{h\in\mathcal{D}_i}\gamma_h^0\gamma_h^{0\prime}\,\mathbb{E}\left(\partial_{z^2}\ell_{ih}\right)\right]^{-1}\gamma_j^0\;\mathbb{E}\left[\partial_z\ell_{ij}\,D_\pi\Delta_{ij} + \frac{1}{2}D_{\pi^2}\Delta_{ij}\right]\right\},$$

$$\overline{D}_\infty^\delta = -\overline{\mathbb{E}}\left\{\frac{1}{J}\sum_{(i,j)\in\mathcal{D}} \alpha_i^{0\prime}\left[\sum_{h\in\mathcal{D}_j}\alpha_h^0\alpha_h^{0\prime}\,\mathbb{E}\left(\partial_{z^2}\ell_{hj}\right)\right]^{-1}\alpha_i^0\;\mathbb{E}\left[\partial_z\ell_{ij}\,D_\pi\Delta_{ij} + \frac{1}{2}D_{\pi^2}\Delta_{ij}\right]\right\},$$

$$\overline{V}_\infty^\delta = -\overline{\mathbb{E}}\left\{\frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left(\Gamma_{ij}\Gamma_{ij}' + \Gamma_{ji}\Gamma_{ij}'\right)\right\},$$

where $\Gamma_{ij} = \overline{(D_\beta\Delta)}_\infty \overline{W}_\infty^{-1}\partial_z\ell_{ij}\tilde{X}_{ij} - \Psi_{ij}\partial_z\ell_{ij}$. *Then,*

$$\sqrt{n}\left[\widehat{\delta} - \delta^0 - \frac{I}{n}\,\overline{(D_\beta\Delta)}_\infty\overline{W}_\infty^{-1}\overline{B}_\infty - \frac{J}{n}\,\overline{(D_\beta\Delta)}_\infty\overline{W}_\infty^{-1}\overline{D}_\infty - \frac{I}{n}\,\overline{B}_\infty^\delta - \frac{J}{n}\,\overline{D}_\infty^\delta\right] \to_d \mathcal{N}(0,\,\overline{V}_\infty^\delta).$$

Remark 2 (Panel Data)**.** *In case (a) of Assumption 1(i), the term involving the cross products* $\Gamma_{ji}\Gamma_{ij}'$ *drops out from the asymptotic variance* $\overline{V}_\infty^\delta$.

Theorem 2 shows that $\widehat{\delta}$ is consistent and normally distributed, but can have bias of the same order as its standard deviation. The first two terms of the bias come from the bias of $\widehat{\beta}$. They drop out when either $\widehat{\beta}$ does not have bias or the APE is estimated from a bias corrected estimator of β. We verify the presence of bias in two of the running examples.

Example 1 (Linear model). *In this case* $\overline{B}_\infty = \overline{D}_\infty = 0$ *and*

$$\Delta_{ij}(\beta,\pi) = (Y_{ij} - X_{ij}'\beta - \pi)^2,$$

so that $\partial_z \Delta_{ij} = -2(Y_{ij} - X_{ij}'\beta^0 - \pi_{ij}^0)$ *and* $\partial_{z^2}\Delta_{ij} = 2$. *Substituting these values in the expressions of the bias of Theorem 2 yields*

$$\overline{B}_\infty^\delta = \overline{D}_\infty^\delta = -R\sigma^2,$$

where we use (4.2). *This result formalizes the analysis in Section 3*

Example 3 (Count response model). *Let* $\Delta_{ij}(\beta,\pi)$ *be as defined in either* (2.5) *or* (2.6). *In this case* $\overline{B}_\infty = \overline{D}_\infty = 0$, *and* $\partial_z \Delta_{ij} = \partial_{z^2}\Delta_{ij} = \Delta_{ij}$. *Substituting these values in the expressions of the bias of Theorem 2 yields*

$$\overline{B}_\infty^\delta = \overline{D}_\infty^\delta = 0,$$

which generalizes the result in Fernández-Val and Weidner (2016) of no asymptotic bias for the estimators of the APEs in the Poisson model with strictly exogenous covariates and additive individual and time effects to the Poisson model with strictly exogenous covariates and multiple factors.

4.3 Bias corrections

Theorems 1 and 2 establish that the estimators of the model parameter and APEs have a bias of the same order as their standard deviations in some models. In this section, we briefly describe how to apply recent developments in nonlinear panel data to correct the bias from the estimators. To simplify the notation we assume that there is no missing data.[4] We consider a generic estimator $\widehat{\theta}$ of the parameter θ, which may correspond to the model parameter or an APE. In this notation, Theorems 1 and 2 show that $\widehat{\theta}$ can have a bias $\mathcal{B}_\infty = \overline{\mathbb{E}}[\mathcal{B}(\beta^0,\phi_n^0)]$ with structure

$$\mathcal{B}(\beta,\phi_n) = \frac{B(\beta,\phi_n)}{J} + \frac{D(\beta,\phi_n)}{I}.$$

The intuition behind this structure is that there are J observations that are informative to estimate each α_i and I observations that are informative to estimate each γ_j.

An analytical correction based on Hahn and Newey (2004) and Fernández-Val and Weidner (2016) can be formed as

$$\widetilde{\theta}_{ABC} = \widehat{\theta} - \widehat{\mathcal{B}}, \quad \widehat{\mathcal{B}} = \mathcal{B}(\widehat{\beta},\widehat{\phi}_n).$$

A split-sample correction based on Dhaene and Jochmans (2015) and Fernández-Val and Weidner (2016) can be formed as

$$\widetilde{\theta}_{SBC} = 3\widehat{\theta} - \bar{\theta}_{I,J/2} - \bar{\theta}_{I/2,J},$$

[4] We refer to Fernández-Val and Weidner (2017) for a discussion on how to modify the corrections to deal with missing data.

where $\bar{\theta}_{I,J/2}$ is the average of the estimators in the haft-panels $\{(i,j) : i = 1, \dots, I; j = 1, \dots, \lceil J/2 \rceil\}$ and $\{(i,j) : i = 1, \dots, I; j = \lfloor J/2 + 1 \rfloor, \dots, J\}$, and $\bar{\theta}_{I/2,J}$ is the average of the estimators in the haft-panels $\{(i,j) : i = 1, \dots, \lceil I/2 \rceil; j = 1, \dots, J\}$ and $\{(i,j) : i = \lfloor I/2 + 1 \rfloor, \dots, I; j = 1, \dots, J\}$, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceil and floor functions. For network data where $I = J$ and the two dimensions of the data index the same entities, Cruz-Gonzalez et al. (2017) proposed the leave-one-out correction

$$\widetilde{\theta}_{NBC} = I\widehat{\theta} - (I-1)\bar{\theta}_{I-1}, \quad \bar{\theta}_{I-1} = I^{-1}\sum_{i=1}^{I} \widehat{\theta}_{-i},$$

where $\widehat{\theta}_{-i}$ is the estimator in the subpanel $\{(k,j) : k = 1, \dots, I; j = 1, \dots, I, k \neq i, j \neq i\}$, that is, the original panel leaving out the observations corresponding to the entity i as either sender or receiver.

As in nonlinear panel data, under suitable conditions the corrections remove the bias without affecting the dispersion. As a result, asymptotic confidence intervals constructed around the corrected estimators are expected to have coverages close to the nominal level in large samples.

5 Numerical Examples

5.1 Gravity Equation with Multiple Latent Factors

The gravity equation is a fundamental empirical relationship in international economics. We estimate a gravity equation of trade between countries using data from Helpman et al. (2008) on bilateral trade flows and other trade-related variables for 157 countries in 1986.[5] The data set contains a network of trade data where both i and j index countries as senders (exporters) and receivers (importers), such that $I = J = 157$. The outcome Y_{ij} is the volume of trade in thousands of constant 2000 US dollars from country i to country j, and the covariates X_{ij} include determinants of bilateral trade flows such as the logarithm of the distance in kilometers between country i's capital and country j's capital and indicators for common colonial ties, currency union, regional free trade area (FTA), border, legal system, language, and religion. Table 3 reports descriptive statistics of the variables used in the analysis. There are $157 \times 156 = 24,492$ observations corresponding to different pairs of countries. The observations with $i = j$ are missing because we do not observe trade flows from a country to itself. The trade variable in the first row is an indicator of positive volume of trade. There are no trade flows for 55% of the country pairs.

[5]The original data set includes 158 countries. We exclude Congo because it did not export to any other country in 1986.

Table 3: Descriptive Statistics

	Mean	Std. Dev.
Trade	0.45	0.50
Trade Volume	84,542	1,082,219
Log Distance	4.18	0.78
Legal	0.37	0.48
Language	0.29	0.45
Religion	0.17	0.25
Border	0.02	0.13
Currency	0.01	0.09
FTA	0.01	0.08
Colony	0.01	0.10
Country Pairs	24,492	

Source: Helpman et al. (2008)

We estimate a Poisson model with the following specification of the intensity

$$\mathbb{E}[Y_{ij} \mid X_{ij}, \alpha_{1i}, \gamma_{1j}, \alpha_{2i}, \gamma_{2j}] = \exp(X_{ij}'\beta + \alpha_{1i} + \gamma_{1j} + \alpha_{2i}'\gamma_{2j}),$$

where α_{2i} and γ_{2i} are R_2-dimensional vectors of factors and factor loadings. This model is a special case of Example 3 with $\alpha_i = (\alpha_{1i}, 1, \alpha_{2i}')'$, $\gamma_j = (1, \gamma_{1j}, \gamma_{2j}')'$, and $R = 2 + R_2$. We explicitly include additive importer and exporter effects to account for scale and multilateral resistance effects following Eaton and Kortum (2001) and Anderson and van Wincoop (2003). Moreover, we also include interactive country effects to capture clustering and homophily induced by latent factors such as country trade partnerships, presence of multinationals or immigrant communities, or differences in natural resources or industrial composition.

Table 4 reports the estimates and standard errors of the parameter β.[6] We consider specifications with different number of interactive effects, R_2, in addition to the additive effects . The last row of the table reports the maximum value of the average log-likelihood, $L(\widehat{\beta}, \widehat{\phi}_n)/n$. In this case, we select $R_2 = 5$ as our preferred specification as adding one more interactive effect only leads to a modest increase in the average log-likelihood, while the precision of some of the parameter estimates starts deteriorating. Moreover, none of the estimates is very sensitive to the choice of R_2 beyond $R_2 = 5$ relative to their standard errors.

[6] We do not report estimates of APEs because in the specification of the Poisson model that we use the parameters can be interpreted as elasticities.

Table 4: Gravity Equation: Parameters

	$R_2 = 0$	$R_2 = 1$	$R_2 = 2$	$R_2 = 3$	$R_2 = 4$	$R_2 = 5$	$R_2 = 6$
Log Distance	-0.64	-0.63	-0.71	-0.69	-0.77	-0.90	-1.01
	(0.04)	(0.05)	(0.05)	(0.07)	(0.08)	(0.10)	(0.11)
Border	0.71	0.41	0.32	0.36	0.38	0.36	0.27
	(0.09)	(0.10)	(0.11)	(0.11)	(0.13)	(0.13)	(0.15)
Legal	0.30	0.14	0.26	0.22	0.13	0.16	0.27
	(0.06)	(0.07)	(0.08)	(0.09)	(0.09)	(0.10)	(0.13)
Language	-0.17	-0.19	-0.02	0.03	-0.09	-0.03	0.09
	(0.10)	(0.12)	(0.12)	(0.13)	(0.14)	(0.15)	(0.17)
Colony	0.36	0.58	0.39	0.45	0.63	0.61	0.55
	(0.13)	(0.15)	(0.15)	(0.17)	(0.19)	(0.19)	(0.22)
Currency	0.60	0.29	1.37	1.38	0.65	0.63	0.77
	(0.48)	(0.71)	(0.68)	(0.68)	(0.80)	(0.83)	(0.96)
FTA	0.25	0.15	0.17	0.13	0.25	0.31	0.26
	(0.11)	(0.12)	(0.13)	(0.14)	(0.16)	(0.17)	(0.18)
Religion	-0.25	0.18	0.24	0.34	0.44	0.30	0.35
	(0.18)	(0.19)	(0.23)	(0.25)	(0.26)	(0.27)	(0.34)
Log-likelihood	-0.44	0.31	0.67	0.84	0.96	1.04	1.11

Notes: all the columns include importer and exporter additive effects.

Standard errors in parenthesis. Log-likelihood is multiplied by 100.

We find that the sign of most of the effects is robust to the inclusion of latent factors. The only exceptions are the effects of common religion and language, which in the specification with only additive effects have counterintuitive negative signs that turn positive after including a sufficient number of factors. Comparing across columns, we observe that the model without factors seems to exaggerate the role of common border, whereas it downplays the effect of distance and colonial links. For example, increasing by 10% the distance reduces by 9% the volume of trade and sharing border increases it by 36% according to our preferred specification with $R_2 = 5$, whereas the same effects are 6% and 71% according to the specification with $R_2 = 0$. The effects of common legal system, currency union and FTA remain positive and statistically significant after the inclusion the factors, without any clear pattern of change. Overall, increasing the number of factors makes the estimates less precise due to the loss of degrees of freedom. This observation showcases a trade-off in estimation between efficiency and robustness to richer dependence structures in the unobservables.

5.2 Calibrated Monte Carlo Simulation

We evaluate the finite-sample properties of our estimation and inference methods in a Monte Carlo simulation that mimics the trade application. The design is calibrated to the Poisson model with additive importer and exporter country effects and one factor. We analyze the performance of the estimator of β in terms of bias, dispersion and inference accuracy. To speed up computation, we include only one covariate: the log distance. More specifically, we generate Y_{ij} from a Poisson distribution with intensity $\exp(X_{ij}\widehat{\beta} + \widehat{\alpha}_{1i} + \widehat{\gamma}_{1j} + \widehat{\alpha}_{2i}\widehat{\gamma}_{2j})$ independently across i and j, where X_{ij} takes the values of log distance in the trade data set, and $\widehat{\beta}$ and $\{\widehat{\alpha}_{1i}, \widehat{\alpha}_{2i}, \widehat{\gamma}_{1i}, \widehat{\gamma}_{2i}\}_{j=1}^{157}$, are equal to the estimates of the parameter, importer effects, exporter effects, factors and factor loadings. We repeat this procedure in 500 simulations for four different sample sizes: $I = 50$, $I = 75$, $I = 100$ and $I = 157$ (full sample in the application). For each sample size and simulation, we draw a random sample of I countries both as importers and exporters without replacement, so that the number of observations is $I \times (I - 1)$. For each simulated sample, we reestimate the model parameter and standard errors, and construct 95% confidence interval for the model parameter.

Table 5 reports the bias (Bias), standard deviation (SD), and root mean squared error (RMSE) of the estimator of the parameter β, together with the ratio of average standard error to the simulation standard deviation (SE/SD), and the empirical coverage in percentage of a confidence interval with 95% nominal value (p;95). We estimate models with three different numbers of factors in addition to the additive effects, $R_2 \in \{1, 2, 3\}$. The results for the bias, sd and rmse are reported in percentage of the true parameter value. We find that the bias is smaller than the standard deviation for every sample size. The confidence intervals cover the parameter in more

Table 5: Results of Calibrated Simulations

I	Bias	SD	RMSE	SE/SD	p;95
			$R_2 = 1$		
50	5.70	14.14	15.22	1.19	98
75	4.63	7.41	8.72	1.16	95
100	1.98	6.45	6.73	1.14	97
157	1.48	3.59	3.87	1.14	96
			$R_2 = 2$		
50	5.69	16.68	17.59	1.08	96
75	4.17	9.96	10.79	1.07	95
100	1.02	7.22	7.28	1.15	96
157	-0.48	4.56	4.58	1.08	96
			$R_2 = 3$		
50	15.25	24.17	28.54	0.95	88
75	4.80	11.55	12.48	1.02	93
100	-0.05	10.17	10.16	1.07	97
157	-0.64	6.59	6.58	1.02	94

Notes: 500 simulations calibrated to trade data.

than 95% of the simulations. The excess coverage is due to the overestimation of the dispersion of the estimators by the standard errors, and decreases with the sample size. Adding unnecessary factors to the specification increases the bias and dispersion of the estimator, but the confidence intervals continue having good coverage properties. This robustness to the inclusion of too many factors is consistent with the theoretical results of Moon and Weidner (2015) for linear factor models. Overall, the simulations show that the asymptotic theory of Section 4 provides a good approximation to the finite-sample behavior of the estimator.

References

Anderson, J. E. and van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Ando, T. and Bai, J. (2016). Large scale panel choice model with unobserved heterogeneity. *Unpublished manuscript.*

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80.

Boneva, L. and Linton, O. (2017). A discrete-choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance. *J. Appl. Econometrics*, 32(7):1226–1243.

Charbonneau, K. (2012). Multiple fixed effects in nonlinear panel data models. *Unpublished manuscript.*

Chen, M. (2014). Estimation of nonlinear panel models with multiple unobserved effects. *Warwick Economics Research Paper Series No. 1120.*

Chen, M., Fernandez-Val, I., and Weidner, M. (2014). Nonlinear Panel Models with Interactive Effects. *ArXiv e-prints.*

Cruz-Gonzalez, M., Fernndez-Val, I., and Weidner, M. (2017). Bias corrections for probit and logit models with two-way fixed effects. *Stata Journal*, 17(3):517–545.

de Paula, A. (2017). Econometrics of network models. In Honore, B., Pakes, A., Piazzesi, M., and Samuelson, L., editors, *Advances in Economics and Econometrics: Theory and Applications: Eleventh World Congress*, Econometric Society Monographs, pages 268–323. Cambridge University Press.

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

Dzemski, A. (2017). An empirical model of dyadic link formation in a network with unobserved heterogeneity. *Unpublished manuscript.*

Eaton, J. and Kortum, S. (2001). Trade in capital goods. *European Economic Review*, 45(7):1195–1235.

Fernández-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics*, 192(1):291–312.

Fernández-Val, I. and Weidner, M. (2017). Fixed Effect Estimation of Large T Panel Data Models. *ArXiv e-prints.*

Graham, B. S. (2015). Methods of identification in social networks. *Annual Review of Economics*, 7(1):465–485.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation. *NBER Working Paper.*

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063.

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

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A*, 170(2):301–354.

Harrigan, J. (1994). Scale economies and the volume of trade. *The Review of Economics and Statistics*, pages 321–328.

Head, K. and Mayer, T. (2014). Chapter 3 - gravity equations: Workhorse,toolkit, and cookbook. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, pages 131 – 195. Elsevier.

Helpman, E., Melitz, M., and Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics*, 123(2):441–487.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.*, 100(469):286–295.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.*, 97(460):1090–1098.

Jochmans, K. (2017). Two-way models for gravity. *Review of Economics and Statistics*, 99(3):478–485.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.

Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.

Moon, H. R. and Weidner, M. (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195.

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

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Santos Silva, J. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4):641–658.

Snijders, T. A. (2011). Statistical models for social networks. *Annual Review of Sociology*, 37(1):131–153.

Yan, T. (2018). Undirected network models with degree heterogeneity and homophily. *ArXiv e-prints*.

Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2016). Statistical inference in a directed network model with covariates. *arXiv preprint arXiv:1609.04558*.

A Proofs

A.1 Notation and Normalization

Remember the log-likelihood defined in the main text, and also define the rescaled version,

$$\mathcal{L}^*(\beta,\phi) := n^{1/2}\, L(\beta,\phi), \qquad\qquad L(\beta,\phi) := \sum_{(i,j)\in\mathcal{D}} \log f(Y_{ij} \mid X_{ij}'\beta + \pi_{ij}).$$

For the true value of the fixed effect parameters $\phi^0 = (\alpha^{0\prime}, \gamma^{0\prime})'$ we impose the normalization $\sum_{i=1}^I \alpha_i^0\, \alpha_i^{0\prime} = \sum_{j=1}^J \gamma_j^0\, \gamma_j^{0\prime}$, and define the restricted parameter set

$$\Phi := \left\{ \phi \in \mathbb{R}^{d_\phi} \;:\; \sum_{i=1}^I \alpha_i^0\, \alpha_i' = \sum_{j=1}^J \gamma_j\, \gamma_j^{0\prime} \right\},$$

where $d_v := \dim v$ for any vector v. Notice that $\phi^0 \in \Phi$. The maximum likelihood estimator that imposes the normalization $\phi \in \Phi$ reads

$$(\widehat{\beta}, \widehat{\phi}) = \operatorname*{argmax}_{(\beta,\phi)\in\mathbb{R}^{d_\beta}\times\Phi} L(\beta,\phi). \tag{A.1}$$

Imposing $\widehat{\phi} \in \Phi$ is an infeasible normalization, because the true value of the parameters appear in the definition of Φ. However, all our final asymptotic results are on the estimators $\widehat{\beta}$ and $\widehat{\delta}$, which are invariant to the chosen normalization for $\widehat{\phi}$, that is, those results on $\widehat{\beta}$ and $\widehat{\delta}$ also hold unchanged for any other normalization, and imposing $\widehat{\phi} \in \Phi$ is simply a matter of convenience for the following proofs.

Notice that in the definition of Φ there are R^2 normalization constraints, and those constraints are linear in ϕ. It is this linearity which makes this normalization so attractive for our purposes. In particular, instead of imposing this normalization directly we can also impose it via a quadratic penalty function by defining the penalized objective function

$$\mathcal{L}(\beta,\phi) = n^{-1/2}\left[L(\beta,\phi) - \frac{b}{2}\phi'\, V\, V'\, \phi \right],$$

where $b > 0$ is some constant, and V is a $d_\phi \times R^2$ matrix, which depends on α^0 and γ^0, and is implicitly defined by

$$V'\,\phi = \operatorname{vec}\left[\sum_{i=1}^I \alpha_i^0 \alpha_i' - \sum_{j=1}^J \gamma_j \gamma_j^{0\,\prime} \right].$$

Thus, the above penalty term can also be expressed as

$$\phi'\, V\, V'\, \phi = \left\| \sum_{i=1}^I \alpha_i^0 \alpha_i' - \sum_{j=1}^J \gamma_j \gamma_j^{0\,\prime} \right\|_F^2,$$

where $\|.\|_F$ denotes the Frobenius norm. The constrained estimator in (A.1) can then equivalently be obtained by solving the unconstrained problem

$$(\widehat{\beta}, \widehat{\phi}) = \underset{(\beta,\phi)\in\mathbb{R}^{d_\beta+d_\phi}}{\operatorname{argmax}} \; \mathcal{L}(\beta,\phi),$$

and we also define

$$\widehat{\phi}(\beta) = \underset{\phi\in\mathbb{R}^{d_\phi}}{\operatorname{argmax}} \; \mathcal{L}(\beta,\phi), \qquad\qquad (\widehat{\alpha}'(\beta), \widehat{\gamma}'(\beta)) = \widehat{\phi}'(\beta).$$

Finally, we introduce the index sets $\mathbf{I} := \{1, \ldots, I\}$ and $\mathbf{J} := \{1, \ldots, J\}$.

A.2 Consistency

Lemma 1. *Let Assumption 1 be satisfied. Then we have* $\|\widehat{\beta} - \beta^0\| = \mathcal{O}_P(I^{-3/8})$ *and*

$$\frac{1}{\sqrt{n}} \left\| \widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime} \right\|_F = \mathcal{O}_P(I^{-3/8} + \|\beta - \beta^0\|), \quad \frac{1}{\sqrt{I}} \|\widehat{\phi}(\beta) - \phi^0\| = \mathcal{O}_P(I^{-3/8} + \|\beta - \beta^0\|),$$

uniformly over β in a ϵ-neighborhood around β^0, for some $\epsilon > 0$.

Proof of Lemma 1. This proof considers $\mathcal{D}_0 = \mathcal{D}$ for simplicity, but since we only allow for a vanishing fraction of missing datapoints $\mathcal{D}_0 \setminus \mathcal{D}$, those cannot affect the consistency and convergence rates of the estimators. For all $z_1, z_2 \in \mathcal{B}^0_\varepsilon$ a second order Taylor expansion of $\ell_{ij}(z_1)$ around z_2 gives

$$\begin{aligned}
\ell_{ij}(z_1) - \ell_{ij}(z_2) &= [\partial_z \ell_{ij}(z_1)](z_1 - z_2) - \tfrac{1}{2}[\partial_{z^2}\ell_{ij}(\tilde{z})]\,(z_1 - z_2)^2 \\
&\geq [\partial_z \ell_{ij}(z_1)](z_1 - z_2) + \frac{b_{\min}}{2}\,(z_1 - z_2)^2 \\
&= \frac{b_{\min}}{2}\left(z_1 - z_2 + \frac{1}{b_{\min}}[\partial_z \ell_{ij}(z_1)]\right)^2 - \frac{1}{2b_{\min}}[\partial_z \ell_{ij}(z_1)]^2. \qquad \text{(A.2)}
\end{aligned}$$

where $\tilde{z} \in [\min(z_1, z_2), \max(z_1, z_2)]$. Let $e_{ij} := \partial_z \ell_{ij}/b_{\min}$. Using (A.2) we find that

$$\begin{aligned}
0 &\geq \frac{1}{\sqrt{IJ}}\left[\mathcal{L}(\beta^0, \phi^0) - \mathcal{L}(\widehat{\beta}, \widehat{\phi})\right] \\
&= \frac{1}{IJ}\sum_{i,j\in\mathcal{D}}\left[\ell_{ij}(z^0_{ij}) - \ell_{ij}(\widehat{z}_{ij})\right] \\
&\geq \frac{b_{\min}}{2\,IJ}\sum_{i,j\in\mathcal{D}}\left[(z^0_{ij} - \widehat{z}_{ij} + e_{ij})^2 - e^2_{ij}\right] \\
&= \frac{b_{\min}}{2\,IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left[(z^0_{ij} - \widehat{z}_{ij} + e_{ij})^2 - e^2_{ij}\right] + \mathcal{O}_P\left(\frac{IJ - n}{IJ}\right) \\
&= \frac{b_{\min}}{2\,IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\left[X'_{ij}(\widehat{\beta} - \beta^0) + \widehat{\alpha}'_i\,\widehat{\gamma}_j - \alpha^{0\prime}_i\,\gamma^0_j - e_{ij}\right]^2 - e^2_{ij}\right\} + \mathcal{O}_P\left(\frac{1}{IJ}\right).
\end{aligned}$$

Note that the penalty term of the objective function does not enter here, because it is zero when evaluated both at the estimates or at the true values of the parameters. Let e be the $I \times J$ matrix with entries e_{ij}. Let X_k be the $I \times J$ matrix with entries $X_{k,ij}$, $k = 1, \ldots, d_\beta$. Let $\beta \cdot X = \sum_k \beta_k X_k$. In matrix notation, the above inequality reads

$$\mathrm{Tr}(e'e) \geq \mathrm{Tr}\left[\left((\widehat{\beta} - \beta^0) \cdot X + \widehat{\alpha}\widehat{\gamma}' - \alpha^0\gamma^{0\prime} - e\right)\left((\widehat{\beta} - \beta^0) \cdot X + \widehat{\alpha}\widehat{\gamma}' - \alpha^0\gamma^{0\prime} - e\right)'\right].$$

Analogous to the consistency proof for linear regression models with interactive fixed effects in Bai (2009) and Moon and Weidner (2017) we can conclude that

$$\begin{aligned}
\frac{1}{IJ}\mathrm{Tr}(e'e) &\geq \frac{1}{IJ}\mathrm{Tr}\left[\mathcal{M}_{\alpha^0}\left((\widehat{\beta} - \beta^0) \cdot X - e\right)\mathcal{M}_{\widehat{\gamma}}\left((\widehat{\beta} - \beta^0) \cdot X - e\right)'\right] \\
&= \frac{1}{IJ}\left[\mathrm{Tr}(e'e) + \mathrm{Tr}\left[\mathcal{M}_{\alpha^0}\left((\widehat{\beta} - \beta^0) \cdot X\right)\mathcal{M}_{\widehat{\gamma}}\left((\widehat{\beta} - \beta^0) \cdot X\right)'\right] + 2\mathrm{Tr}\left[\left((\widehat{\beta} - \beta^0) \cdot X\right)' e\right]\right. \\
&\qquad \left. + \mathcal{O}_P(\|e\|^2) + \mathcal{O}_P(\|\widehat{\beta} - \beta^0\|\|e\| \max_k \|X_k\|)\right] + \mathcal{O}_P\left(\frac{1}{IJ}\right), \qquad \text{(A.3)}
\end{aligned}$$

where we used that e.g.

$$\begin{aligned}
\left|\mathrm{Tr}\left(X_k'\mathcal{P}_{\alpha^0}e\right)\right| &\leq \mathrm{rank}\left(X_k'\mathcal{P}_{\alpha^0}e\right)\left\|X_k'\mathcal{P}_{\alpha^0}e\right\| \leq \|X_k\|\|e\|, \\
\left|\mathrm{Tr}\left(e'\mathcal{P}_{\alpha^0}e\right)\right| &\leq \mathrm{rank}\left(e'\mathcal{P}_{\alpha^0}e\right)\left\|e'\mathcal{P}_{\alpha^0}e\right\| \leq \|e\|^2.
\end{aligned}$$

Lemma D.6 in Fernández-Val and Weidner (2016) shows that under our assumptions $\|\partial_z \ell\| = \mathcal{O}_P(I^{5/8})$, where $\partial_z \ell$ is the $I \times J$ matrix with entries $\partial_z \ell_{ij}$. We thus also have $\|e\| = \mathcal{O}_P(I^{5/8})$. We furthermore have $\|X_k\|^2 \leq \|X_k\|_F^2 = \sum_{ij} X_{k,ij}^2 = \mathcal{O}_P(IJ)$, so that $\|X_k\| = \mathcal{O}_P(\sqrt{IJ})$. We thus have $\|X_k\|\|e\| = \mathcal{O}_P(I^{13/8})$, $\|e\|^2 = \mathcal{O}_P(I^{5/4})$, and

$$\mathrm{Tr}\left(X_k'e\right) = \frac{1}{b_{\min}}\sum_{ij} X_{ij}\partial_z \ell_{ij} = \mathcal{O}_P(\sqrt{IJ}).$$

Applying these results and the generalized collinearity assumption to (A.3) gives

$$0 \geq c\|\widehat{\beta} - \beta^0\| + \mathcal{O}_P(I^{-3/8}\|\widehat{\beta} - \beta^0\|) + \mathcal{O}_P(I^{-3/4}).$$

This implies that $\|\widehat{\beta} - \beta^0\| = \mathcal{O}_P(I^{-3/8})$.

Define $e_{ij}(\beta) = \partial_z \ell_{ij}(X_{ij}'\beta + \alpha_i^0\gamma_j^{0\prime})/b_{\min}$. Analogous to the above argument we find from $\mathcal{L}(\beta, \widehat{\phi}(\beta)) \geq \mathcal{L}(\beta, \phi^0)$ that

$$\begin{aligned}
0 &\geq \sqrt{IJ}\left[\mathcal{L}(\beta, \phi^0) - \mathcal{L}(\beta, \widehat{\phi}(\beta))\right] \\
&= \sum_{i,j}\left[\ell_{ij}(X_{ij}'\beta + \alpha_i^0\gamma_j^{0\prime}) - \ell_{ij}(X_{ij}'\beta + \widehat{\alpha}_i(\beta)\widehat{\gamma}_j'(\beta))\right] \\
&= \frac{b_{\min}}{2}\sum_{i,j}\left\{\left[\widehat{\alpha}_i(\beta)\widehat{\gamma}_j'(\beta) - \alpha_i^0\gamma_j^{0\prime} - e_{ij}(\beta)\right]^2 - [e_{ij}(\beta)]^2\right\}.
\end{aligned}$$

This implies that

$$
\begin{aligned}
&\operatorname{Tr}(e(\beta)'e(\beta)) \\
&\geq \operatorname{Tr}\left[\left(\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime} - e(\beta)\right)\left(\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime} - e(\beta)\right)'\right] \\
&= \operatorname{Tr}(e(\beta)'e(\beta)) + \underbrace{\operatorname{Tr}\left[\left(\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right)\left(\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right)'\right]}_{=\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)'-\alpha^0\gamma^{0\prime}\|_F^2} + \mathcal{O}_P\left(\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F \|e(\beta)\|\right).
\end{aligned}
$$

Since $\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}$ is at most of rank $2R$ we have $\frac{1}{\sqrt{2R}}\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F \leq \left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\| \leq \left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F$, i.e. the Frobenius and the spectral norm are equivalent. Since $e_{ij}(\beta) = e_{ij} + [X_{ij}'(\beta - \beta^0)]\partial_{z^2}\ell_{ij}(X_{ij}'\tilde{\beta} + \alpha_i^0\gamma_j^{0\prime})/b_{\min}$, where $\tilde{\beta}$ lies between β and β^0, we have $\|e(\beta)\| \leq \|e\| + \mathcal{O}_P(\sqrt{IJ}\|\beta - \beta^0\|)$. We thus find

$$
0 \geq \frac{1}{IJ}\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F^2 + \mathcal{O}_P\left[(I^{-3/8} + \|\beta - \beta^0\|)\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F/\sqrt{IJ}\right].
$$

From this we conclude that

$$
\frac{1}{\sqrt{IJ}}\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F = \mathcal{O}_P(I^{-3/8} + \|\beta - \beta^0\|). \tag{A.4}
$$

Next, using our normalization $\widehat{\phi} \in \Phi$ we have

$$
\alpha^{0\prime}\left[\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right]\gamma^0 = \left[\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^2 - \left[\alpha^{0\prime}\alpha^0\right]^2,
$$

and therefore

$$
\begin{aligned}
\left\|\left[\frac{1}{I}\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^2 - \left[\frac{1}{I}\alpha^{0\prime}\alpha^0\right]^2\right\|_F &= \frac{1}{I^2}\left\|\alpha^{0\prime}\left[\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right]\gamma^0\right\|_F \leq \frac{1}{I^2}\left\|\alpha^0\right\|_F\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right\|_F\left\|\gamma^0\right\| \\
&= \frac{1}{I^2}\mathcal{O}(I^{1/2})\sqrt{IJ}\,\mathcal{O}_P(I^{-3/8} + \|\beta - \beta^0\|)\mathcal{O}(J^{1/2}) = \mathcal{O}_P(I^{-3/8} + \|\beta - \beta^0\|).
\end{aligned}
$$

Using the strong-factor assumption $I^{-1}\alpha^{0\prime}\alpha^0 \to_P \Sigma_1 > 0$ we thus have

$$
\left[I^{-1}\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^{-1} = \left[I^{-1}\alpha^{0\prime}\alpha^0\right]^{-1} + \mathcal{O}_P(I^{-3/8} + \|\beta - \beta^0\|). \tag{A.5}
$$

Again by the normalization $\widehat{\phi} \in \Phi$ we also have

$$
\left[\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right]\gamma^0 = \widehat{\alpha}(\beta)\alpha^{0\prime}\widehat{\alpha}(\beta) - \alpha^0\alpha^{0\prime}\alpha^0,
$$

and therefore

$$
\widehat{\alpha}(\beta) = \alpha^0\left[I^{-1}\alpha^{0\prime}\alpha^0\right]\left[I^{-1}\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^{-1} - I\left[\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)' - \alpha^0\gamma^{0\prime}\right]\gamma^0\left[I^{-1}\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^{-1}.
$$

Applying (A.4) and (A.5) thus gives

$$\begin{aligned}
I^{-1/2}\left\|\widehat{\alpha}(\beta)-\alpha^0\right\|_F &\leq I^{-1/2}\left\|\alpha^0\right\|_F\left\|\mathbb{I}_R-\left[I^{-1}\alpha^{0\prime}\alpha^0\right]\left[I^{-1}\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^{-1}\right\|_F \\
&\quad + I^{-3/2}\left\|\widehat{\alpha}(\beta)\widehat{\gamma}(\beta)'-\alpha^0\gamma^{0\prime}\right\|_F\left\|\gamma^0\right\|_F\left\|\left[I^{-1}\alpha^{0\prime}\widehat{\alpha}(\beta)\right]^{-1}\right\|_F \\
&= I^{-1/2}\mathcal{O}(I^{1/2})\mathcal{O}_P(I^{-3/8}+\|\beta-\beta^0\|)+I^{-3/2}\sqrt{IJ}\,\mathcal{O}_P(I^{-3/8}+\|\beta-\beta^0\|)\mathcal{O}(J^{1/2})\mathcal{O}(1) \\
&= \mathcal{O}_P(I^{-3/8}+\|\beta-\beta^0\|).
\end{aligned}$$

Analogously we conclude that $J^{-1/2}\|\widehat{\gamma}(\beta)-\gamma^0\|=\mathcal{O}_P(I^{-3/8}+\|\beta-\beta^0\|)$, and therefore $\frac{1}{\sqrt{I}}\|\widehat{\phi}(\beta)-\phi^0\|=\mathcal{O}_P(I^{-3/8}+\|\beta-\beta^0\|)$. ■

A.3 Inverse Expected Incidental Parameter Hessian

We define the expected incidental parameter Hessian for the log-likelihood with and without the penalty term as

$$\overline{\mathcal{H}}^*=\mathbb{E}[-\partial_{\phi\phi'}\mathcal{L}^*], \qquad \overline{\mathcal{H}}=\mathbb{E}[-\partial_{\phi\phi'}\mathcal{L}]=\overline{\mathcal{H}}^*+\frac{b}{\sqrt{n}}VV',$$

Let $a=\operatorname{vec}(\alpha)$ and $c=\operatorname{vec}(\gamma)$, so that $\phi=(a',c')'$. Correspondingly we can decompose the Hessian matrix,

$$\overline{\mathcal{H}}^*=\begin{pmatrix}\mathbb{E}[-\partial_{aa'}\mathcal{L}^*] & \mathbb{E}[-\partial_{ac'}\mathcal{L}^*] \\ \mathbb{E}[-\partial_{ca'}\mathcal{L}^*] & \mathbb{E}[-\partial_{cc'}\mathcal{L}^*]\end{pmatrix}=:\begin{pmatrix}\overline{\mathcal{H}}^*_{(\alpha\alpha)} & \overline{\mathcal{H}}^*_{(\alpha\gamma)} \\ [\overline{\mathcal{H}}^*_{(\alpha\gamma)}]' & \overline{\mathcal{H}}^*_{(\gamma\gamma)}\end{pmatrix}.$$

Here, $\overline{\mathcal{H}}^*_{(\alpha\alpha)}$ is a block-diagonal $IR\times IR$ matrix with $R\times R$ diagonal blocks, and $\overline{\mathcal{H}}^*_{(\gamma\gamma)}$ is a block-diagonal $JR\times JR$ matrix with $R\times R$ diagonal blocks, which themselves are composed of $R\times R$ blocks of the form

$$\overline{\mathcal{H}}^*_{(\alpha\alpha)}=\operatorname{diag}\left(\left[\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{D}_i}\mathbb{E}(-\partial_{z^2}\ell_{ij})\gamma_j^0\gamma_j^{0\prime}\right]_{i\in\mathbf{I}}\right), \quad \overline{\mathcal{H}}^*_{(\gamma\gamma)}=\operatorname{diag}\left(\left[\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{D}_j}\mathbb{E}(-\partial_{z^2}\ell_{ij})\alpha_j^0\alpha_j^{0\prime}\right]_{j\in\mathbf{J}}\right).$$

For any matrix A with elements A_{kl} let $\|A\|_{\max}=\max_{k,l}|A_{kl}|$. Notice that $\|.\|_{\max}$ is not sub-multiplicative, so it is not a matrix norm.

Lemma 2. *Under Assumptions 1 we have*

$$\left\|\overline{\mathcal{H}}^{-1}-\operatorname{diag}\left(\overline{\mathcal{H}}^*_{(\alpha\alpha)},\overline{\mathcal{H}}^*_{(\gamma\gamma)}\right)^{-1}\right\|_{\max}=\mathcal{O}\left(n^{-1/2}\right).$$

Proof. We consider the case $\mathcal{D}=\mathcal{D}_0$ in the following. We decompose

$$\overline{\mathcal{H}}^*=\underbrace{\begin{pmatrix}\overline{\mathcal{H}}^*_{(\alpha\alpha)} & 0 \\ 0 & \overline{\mathcal{H}}^*_{(\gamma\gamma)}\end{pmatrix}}_{=:\overline{\mathcal{D}}}+\underbrace{\begin{pmatrix}0 & \overline{\mathcal{H}}^*_{(\alpha\gamma)} \\ [\overline{\mathcal{H}}^*_{(\alpha\gamma)}]' & 0\end{pmatrix}}_{=:\overline{\mathcal{A}}^*},$$

and let $\overline{\mathcal{A}} := \overline{\mathcal{A}}^* + \frac{b}{\sqrt{n}} VV'$. We then have $\overline{\mathcal{H}} = \overline{\mathcal{D}} + \overline{\mathcal{A}}$. The $IR \times JR$ matrix $\overline{\mathcal{H}}^*_{(\alpha\gamma)}$ is composed of $I \times J$ blocks of size $R \times R$ as follows

$$\overline{\mathcal{H}}^*_{(\alpha\gamma)} = \left[\frac{1}{\sqrt{n}} \mathbb{E}(-\partial_{z^2}\ell_{ij}) \gamma^0_j \alpha^{0\prime}_i \right]_{i\in\mathbf{I}, j\in\mathbf{J}},$$

and similarly we have blocks for the $(I+J)R \times (I+J)R$ matrix VV'

$$VV' = \begin{pmatrix} \left[\alpha^0_i \alpha^{0\prime}_{i^*}\right]_{i,i^*\in\mathbf{I}} & \left[-\gamma^0_j \alpha^{0\prime}_i\right]_{i\in\mathbf{I}, j\in\mathbf{J}} \\ \left[-\alpha^0_i \gamma^{0\prime}_j\right]_{j\in\mathbf{J}, i\in\mathbf{I}} & \left[\gamma^0_j \gamma^{0\prime}_{j^*}\right]_{j,j^*\in\mathbf{J}} \end{pmatrix} =: \begin{pmatrix} [VV']_{(\alpha\alpha)} & [VV']_{(\alpha\gamma)} \\ [VV']_{(\gamma\alpha)} & [VV']_{(\gamma\gamma)} \end{pmatrix}.$$

Let $b^* := \min\{b_{\min}, b\}$. For symmetric matrices A and B we write $A \geq B$ if $A - B$ is positive semi-definite. We have

$$\overline{\mathcal{A}} - \frac{b-b^*}{\sqrt{n}} VV' - \frac{b^*}{\sqrt{n}} \begin{pmatrix} [VV']_{(\alpha\alpha)} & 0 \\ 0 & [VV']_{(\gamma\gamma)} \end{pmatrix} = \begin{pmatrix} 0 & \overline{\mathcal{H}}^*_{(\alpha\gamma)} - b^* [VV']_{(\alpha\gamma)} \\ \left[\overline{\mathcal{H}}^*_{(\alpha\gamma)}\right]' - b^* [VV']_{(\gamma\alpha)} & 0 \end{pmatrix},$$

and since $V'V \geq 0$ (implying also $[VV']_{(\alpha\alpha)} \geq 0$ and $[VV']_{(\gamma\gamma)} \geq 0$) we thus have

$$\overline{\mathcal{A}} \geq \begin{pmatrix} 0 & \overline{\mathcal{H}}^*_{(\alpha\gamma)} - b^* [VV']_{(\alpha\gamma)} \\ \left[\overline{\mathcal{H}}^*_{(\alpha\gamma)}\right]' - b^* [VV']_{(\gamma\alpha)} & 0 \end{pmatrix}.$$

Using this and $\mathbb{E}[-\partial_{\phi\phi'}\ell_{ij}] \geq 0$ we obtain

$$\begin{aligned}
\overline{\mathcal{H}} &= \overline{\mathcal{D}} + \overline{\mathcal{A}} \\
&\geq \overline{\mathcal{D}} - \begin{pmatrix} 0 & \overline{\mathcal{H}}^*_{(\alpha\gamma)} - b^* [VV']_{(\alpha\gamma)} \\ \left[\overline{\mathcal{H}}^*_{(\alpha\gamma)}\right]' - b^* [VV']_{(\gamma\alpha)} & 0 \end{pmatrix} - \underbrace{n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \mathbb{E}[-\partial_{\phi\phi'}\ell_{ij}] \frac{\mathbb{E}(-\partial_{z^2}\ell_{ij}) - b^*}{\mathbb{E}(-\partial_{z^2}\ell_{ij})}}_{\geq 0} \\
&= b^* \begin{pmatrix} \operatorname{diag}\left(\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{I} \gamma^0_i \gamma^{0\prime}_i \right]_{j\in\mathbf{J}} \right) & 0 \\ 0 & \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{J} \alpha^0_j \alpha^{0\prime}_j \right]_{j\in\mathbf{J}} \end{pmatrix} \\
&= b^* \begin{pmatrix} n^{-1/2}\, \mathbb{I}_I \otimes \gamma^{0\prime}\gamma^0 & 0 \\ 0 & n^{-1/2}\, \mathbb{I}_J \otimes \alpha^{0\prime}\alpha^0 \end{pmatrix} \geq c\, \mathbb{I}_{(I+J)R},
\end{aligned}$$

wpa1, where existence of $c > 0$ is guaranteed by our strong factor Assumptions 1(v). The result of the last display implies

$$\overline{\mathcal{H}}^{-1} \leq c^{-1}\, \mathbb{I}_{(I+J)R}. \tag{A.6}$$

We have thus obtained a spectral bound for $\overline{\mathcal{H}}^{-1}$. This turns out to be the key step in the proof. The remainder of the proof is just a relatively straightforward expansion of $\overline{\mathcal{H}}^{-1}$. Namely, using

$\overline{\mathcal{H}} = \overline{\mathcal{D}} + \overline{\mathcal{A}}$ we find that

$$
\begin{aligned}
\overline{\mathcal{H}}^{-1} &= \overline{\mathcal{D}}^{-1} - \overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{D}}^{-1} + \left[\overline{\mathcal{D}}^{-1}\,\overline{\mathcal{H}}\,\overline{\mathcal{D}}^{-1} - 2\overline{\mathcal{D}}^{-1} + \overline{\mathcal{H}}^{-1}\right] \\
&= \overline{\mathcal{D}}^{-1} - \overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{D}}^{-1} + \overline{\mathcal{D}}^{-1}\left(\overline{\mathcal{H}} - \overline{\mathcal{D}}\right)\overline{\mathcal{H}}^{-1}\left(\overline{\mathcal{H}} - \overline{\mathcal{D}}\right)\overline{\mathcal{D}}^{-1} \\
&= \overline{\mathcal{D}}^{-1} - \overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{D}}^{-1} + \overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{H}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{D}}^{-1} \\
&\leq \overline{\mathcal{D}}^{-1} - \overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{D}}^{-1} + c^{-1}\overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}^{2}\overline{\mathcal{D}}^{-1},
\end{aligned}
$$

and therefore

$$
\left\|\overline{\mathcal{H}}^{-1} - \overline{\mathcal{D}}^{-1}\right\|_{\max} \leq \left\|\overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}\,\overline{\mathcal{D}}^{-1}\right\|_{\max} + c^{-1}\left\|\overline{\mathcal{D}}^{-1}\,\overline{\mathcal{A}}^{2}\overline{\mathcal{D}}^{-1}\right\|_{\max}.
$$

From the expressions for $\overline{\mathcal{D}}$ and $\overline{\mathcal{A}}$ above one finds that $\overline{\mathcal{D}}$ is block-diagonal with entries of order one, and $\|\overline{\mathcal{A}}\|_{\max} = \mathcal{O}(n^{-1/2})$, which implies $\left\|\overline{\mathcal{A}}^{2}\right\|_{\max} = \mathcal{O}((I+J)n^{-1}) = \mathcal{O}(n^{-1/2})$. The rhs of the last display is therefore indeed of order $n^{-1/2}$. ■

A.4 Local Concavity of the Objective Function

The consistency results for $\widehat{\beta}$ and $\widehat{\phi}(\beta)$ in Lemma 1 provide some initial convergence rates, implying that we only need to consider a shrinking neighborhood around β^0 and ϕ^0 for the remaining asymptotic analysis. The following lemma shows that the objective function $\mathcal{L}(\beta, \phi)$ is strictly concave in such a local neighborhood. Later in the proof this strict concavity will allow us to apply the general expansion results in Fernández-Val and Weidner (2016).

Lemma 3. *Let Assumption 1 be satisfied, and let $r_\beta = r_{\beta,n} = o_P(1)$ and $r_\phi = r_{\phi,n} = o_P(n^{1/4})$. Then, $\mathcal{H}(\beta, \phi)$ is positive definite for all $\beta \in \mathcal{B}(r_\beta, \beta^0)$ and $\phi \in \mathcal{B}(r_\phi, \phi^0)$, wpa1, where $\mathcal{B}(r_\beta, \beta^0)$ and $\mathcal{B}(r_\phi, \phi^0)$ are balls under the Euclidian norm. This implies that $\mathcal{L}(\beta, \phi)$ is strictly concave in $\phi \in \mathcal{B}(r_\phi, \phi^0)$, for all $\beta \in \mathcal{B}(r_\beta, \beta^0)$.*

Proof. Analogously to the expected incidental parameter Hessian $\overline{\mathcal{H}}$ at the true parameters that was discussed above we now introduce the following notation for incidental parameter Hessian (without expectations, and not necessarily at the true parameters),

$$
\mathcal{H}(\beta, \phi) = -\partial_{\phi\phi'}\mathcal{L}(\beta, \phi) = \begin{pmatrix} \mathcal{H}^*_{(\alpha\alpha)}(\beta, \phi) & \mathcal{H}^*_{(\alpha\gamma)}(\beta, \phi) \\ [\mathcal{H}^*_{(\alpha\gamma)}(\beta, \phi)]' & \mathcal{H}^*_{(\gamma\gamma)}(\beta, \phi) \end{pmatrix} + \frac{b}{\sqrt{n}} VV'.
$$

Let $\ell_{ij}(\beta, \pi_{ij}) := \ell_{ij}(z_{ij})$, where $\pi_{ij} = \alpha_i'\gamma_j$ and $z_{ij} = X_{ij}'\beta + \alpha_i'\gamma_j$. We then have

$$\mathcal{H}^*_{(\alpha\alpha)}(\beta,\phi) = \operatorname{diag}\left(\left[\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{D}_i}[-\partial_{z^2}\ell_{ij}(\beta,\pi_{ij})]\gamma_j^0\gamma_j^{0\prime}\right]_{i\in\mathbf{I}}\right),$$
$$\mathcal{H}^*_{(\gamma\gamma)}(\beta,\phi) = \operatorname{diag}\left(\left[\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{D}_j}[-\partial_{z^2}\ell_{ij}(\beta,\pi_{ij})]\alpha_j^0\alpha_j^{0\prime}\right]_{j\in\mathbf{J}}\right),$$
$$\mathcal{H}^*_{(\alpha\gamma)}(\beta,\phi) = \left\{\frac{1}{\sqrt{n}}[-\partial_{z^2}\ell_{ij}(\beta,\pi_{ij})]\gamma_j^0\alpha_i^{0\prime} + \frac{1}{\sqrt{n}}[-\partial_z\ell_{ij}(z_{ij})]\,\mathbb{I}_R\right\}_{i\in\mathbf{I},j\in\mathbf{J}},$$

We decompose the Hessian into the contribution from the first and from the second derivative of the log-likelihood, namely $\mathcal{H}(\beta,\phi) = H(\beta,\phi) + F(\beta,\phi)$, where

$$F(\beta,\phi) = \begin{pmatrix} 0_{N\times N} & F_{(\alpha\gamma)}(\beta,\phi) \\ [F_{(\alpha\gamma)}(\beta,\phi)]' & 0_{T\times T}\end{pmatrix}, \qquad F_{(\alpha\gamma)}(\beta,\phi) = \left\{\frac{1}{\sqrt{n}}[-\partial_z\ell_{ij}(z_{ij})]\,\mathbb{I}_R\right\}_{i\in\mathbf{I},j\in\mathbf{J}}.$$

Notice that $H(\beta,\phi)$ has the same structure as $\overline{H}$. Analogously to the bound (A.7) derived in the proof of Lemma 2 we can thus show that there exists a constant $c>0$ such that wpa1 we have, for $\phi\in\mathcal{B}(r_\phi,\phi^0)$ and $\beta\in\mathcal{B}(r_\beta,\beta^0)$,

$$H(\beta,\phi) \geq c\,\mathbb{I}_{(I+J)R}. \tag{A.7}$$

The new terms that need to be accounted for here are the first derivative terms $F(\beta,\phi)$, which are zero in expectation at the true parameter and therefore did not show up in our discussion of $\overline{\mathcal{H}}$ above. The goal in the following is to show that $\|F_{(\alpha\gamma)}(\beta,\phi)\| = o_P(1)$, or equivalently $\|F_{(\alpha\gamma)}(\beta,\phi)\| = o_P(1)$, within the shrinking neighborhood of the true parameters. Here, $\|.\|$ refers to the spectral norm.

For ease of notation we consider $R=1$ in the remainder of this proof. We then have $F_{(\alpha\gamma)ij}(\beta,\phi) = -\frac{1}{\sqrt{n}}\partial_\pi\ell_{ij}(\beta,\alpha_i'\gamma_j)$. A Taylor expansion gives

$$\partial_\pi\ell_{ij}(\beta,\alpha_i'\gamma_j) = \partial_\pi\ell_{ij}(\beta^0,\alpha_i^0\gamma_j^0{}') + (\beta-\beta^0)'\partial_{\beta\pi}\ell_{ij}(\tilde\beta_{ij},\tilde\pi_{ij}) + (\alpha_i'\gamma_j - \alpha_i^0\gamma_j^0{}')\partial_{\pi^2}\ell_{ij}(\tilde\beta_{ij},\tilde\pi_{ij}).$$

The spectral norm of the $I\times J$ matrix with entries $\partial_{\beta_k\pi}\ell_{ij}(\tilde\beta_{ij},\tilde\pi_{ij})$ is bounded by the Frobenius norm of this matrix, which is of order $\sqrt{n}$, since we assume uniformly bounded moments for $\partial_{\beta_k\pi}\ell_{ij}(\tilde\beta_{ij},\tilde\pi_{ij})$. The spectral norm of the $I\times J$ matrix with entries $(\alpha_i'\gamma_j-\alpha_i^0\gamma_j^0{}')\partial_{\pi^2}\ell_{ij}(\tilde\beta_{ij},\tilde\pi_{ij})$ is also bounded by the Frobenius norm of this matrix, which equals $\sqrt{\sum_{ij}(\alpha_i'\gamma_j-\alpha_i^0\gamma_j^0{}')^2[\partial_{\pi^2}\ell_{ij}(\tilde\beta_{ij},\tilde\pi_{ij})]^2}$ and thus bounded by $b_{\max}\sqrt{\sum_{ij}(\alpha_i'\gamma_j-\alpha_i^0\gamma_j^0{}')^2} = b_{\max}\|\alpha\gamma'-\alpha^0\gamma^{0\prime}\|_F$. We thus find

$$\begin{aligned}\left\|F_{(\alpha\gamma)ij}(\beta,\phi)\right\| &\leq \frac{1}{\sqrt{n}}\left(\|\partial_\pi\ell_{ij}\| + \mathcal{O}_P(\sqrt{n})\|\beta-\beta^0\| + b_{\max}\|\alpha\gamma'-\alpha^0\gamma^{0\prime}\|_F\right)\\ &= \mathcal{O}_P(\frac{1}{\sqrt{n}}I^{5/8}) + \mathcal{O}_P(r_\beta) + \mathcal{O}_P(r_\phi/\sqrt{I})\\ &= o_P(1),\end{aligned}$$

for $\phi \in \mathcal{B}(r_\phi, \phi^0)$ and $\beta \in \mathcal{B}(r_\beta, \beta^0)$, where we also used that $\|\alpha\gamma' - \alpha^0\gamma^{0\prime}\|_F = \mathcal{O}_P(\sqrt{I})\|\phi - \phi^0\|$.

Combining the result in the last display with (A.7) we find that there exists a constant $c > 0$ such that wpa1 we have, for $\phi \in \mathcal{B}(r_\phi, \phi^0)$ and $\beta \in \mathcal{B}(r_\beta, \beta^0)$,

$$\mathcal{H}(\beta, \phi) \geq c\, \mathbb{I}_{(I+J)R}.$$

We have thus shown that $\mathcal{L}(\beta, \phi)$ is indeed strictly concave (or that $-\mathcal{L}(\beta, \phi)$ is strictly convex) within this shrinking neighborhood. ■

A.5 Stochastic Expansion

Once we have the consistency result of Lemma 1 and the local strict concavity result of Lemma 3, then the remaining stochastic expansion of the fixed effect estimators $\widehat{\beta}$ and $\widehat{\delta}$ does not actually depend on the specific single index and interactive fixed effect structure of our model, and some of the conceptual issues indeed become more transparent when ignoring that structure. Therefore, in this subsection, let $\ell_{ij}(\beta, \alpha_i, \gamma_j) := \ell_{ij}(X_{ij}'\beta + \alpha_i'\gamma_j)$ and $\Delta_{ij}(\beta, \alpha_i, \gamma_j) := \Delta_{ij}(\beta, \pi_{ij})$. Remember that our fixed effect estimators $\widehat{\beta}$ and $\widehat{\gamma}$ maximize the objective function $\mathcal{L}(\beta, \phi) = n^{-1/2}\left[\sum_{(i,j)\in\mathcal{D}} \ell_{ij}(\beta, \alpha_i, \gamma_j) + \frac{b}{2}\phi' V V' \phi\right]$, where $\phi = [(\alpha_i')_{i\in\mathbf{I}}, (\gamma_j')_{j\in\mathbf{J}}]'$. The APE is $\delta^0 = \Delta(\beta^0, \phi^0) = \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \Delta_{ij}(\beta^0, \alpha_i^0, \gamma_j^0)$, and the corresponding plug-in estimator reads $\widehat{\delta} = \Delta(\widehat{\beta}, \widehat{\phi})$. For partial derivatives of $\ell_{ij}(\beta, \alpha_i, \gamma_j)$ and $\Delta(\widehat{\beta}, \widehat{\phi})$ we use superscipts in the following, expectations are always conditional on ϕ and are indicated by a bar, and arguments are omitted when evaluated at the true parameters. For example, $\bar{\ell}_{ij}^{\alpha_i\alpha_i}$ is the $d_\alpha \times d_\alpha$ expected Hessian matrix of $\ell_{ij}(\beta, \alpha_i, \gamma_j)$ evaluated at the true parameters. This is the notation also used in Fernández-Val and Weidner (2017), but here the α_i and γ_j are vectors of length d_α and d_γ, respectively. For our interactive fixed effect model we have $d_\alpha = d_\gamma = R$, but this is not used in the rest of this subsection. The advantage of this generality is that, for example, the following formulas are also applicable to models where in addition to the interactive effects we would include another additive effect in the single index.

It is convenient to make the log-likelihood information-orthogonal between β and the incidental parameters. This can be achieved by the transformation[7]

$$\begin{aligned}
\ell_{ij}^*(\beta, \alpha_i, \gamma_j) &:= \ell_{ij}(\beta, \alpha_i + \xi_i^{(\alpha)}\beta, \gamma_j + \xi_j^{(\gamma)}\beta), \\
\Delta_{ij}^*(\beta, \alpha_i, \gamma_j) &:= \Delta_{ij}(\beta, \alpha_i + \xi_i^{(\alpha)}\beta, \gamma_j + \xi_j^{(\gamma)}\beta),
\end{aligned}$$

[7] This transformation corresponds to the reparameterization $\alpha_i^* = \alpha_i - \xi_i^{(\alpha)}\beta$ and $\gamma_j^* = \gamma_j - \xi_j^{(\gamma)}\beta$. The log-likelihood with respect to these parameters is $\ell_{ij}(\beta, \alpha_i^* + \xi_i^{(\alpha)}\beta, \gamma_j^* + \xi_j^{(\gamma)}\beta) =: \ell_{ij}^*(\beta, \alpha_i^*, \gamma_j^*)$, which gives our definition of ℓ_{ij}^* after renaming (α_i^*, γ_j^*) as (α_i, γ_j) again.

where the $d_\alpha \times d_\beta$ matrices $\xi_i^{(\alpha)}$, and the $d_\gamma \times d_\beta$ matrices ρ_j are a solution to the system of equations

$$\begin{aligned}
\sum_{j\in\mathcal{D}_i} \left[\overline{\ell}_{ij}^{\alpha_i\beta} + \overline{\ell}_{ij}^{\alpha_i\alpha_i} \xi_i^{(\alpha)} + \overline{\ell}_{ij}^{\alpha_i'\gamma_j} \xi_j^{(\gamma)} \right] &= 0, \qquad i = 1,\ldots,I, \\
\sum_{i\in\mathcal{D}_j} \left[\overline{\ell}_{ij}^{\gamma_j\beta} + \overline{\ell}_{ij}^{\gamma_j\alpha_i} \xi_i^{(\alpha)} + \overline{\ell}_{ij}^{\gamma_j\gamma_j} \xi_j^{(\gamma)} \right] &= 0, \qquad j = 1,\ldots,J.
\end{aligned}$$

Analogously, let the d_α-vectors $\psi_i^{(\alpha)}$ and the d_γ-vectors $\psi_j^{(\gamma)}$ be solutions to the system of equations

$$\begin{aligned}
\sum_{j\in\mathcal{D}_i} \left[\overline{\Delta}_{ij}^{\alpha_i} + \overline{\ell}_{ij}^{\alpha_i\alpha_i} \psi_i^{(\alpha)} + \overline{\ell}_{ij}^{\alpha_i'\gamma_j} \psi_j^{(\gamma)} \right] &= 0, \qquad i = 1,\ldots,I, \\
\sum_{i\in\mathcal{D}_j} \left[\overline{\Delta}_{ij}^{\gamma_j} + \overline{\ell}_{ij}^{\gamma_j\alpha_i} \psi_i^{(\alpha)} + \overline{\ell}_{ij}^{\gamma_j\gamma_j} \psi_j^{(\gamma)} \right] &= 0, \qquad j = 1,\ldots,J.
\end{aligned}$$

Finally, let

$$\overline{W} = -\frac{1}{\sqrt{n}} \left(\overline{\mathcal{L}}^{\beta\beta} + \overline{\mathcal{L}}^{\beta\phi}\, \overline{\mathcal{H}}^{-1}\, \overline{\mathcal{L}}^{\phi\beta} \right) = -\frac{1}{\sqrt{n}} \overline{\mathcal{L}}^{*\,\beta\beta} = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \overline{\ell}_{ij}^{*\,\beta\beta}.$$

The $d_\beta \times d_\beta$ matrix $\overline{W}_\infty$ defined in Assumption (1) is simply the probability limit of $\overline{W}$, that is, $\overline{W}_\infty = \overline{\mathbb{E}}\, \overline{W}$ in main text notation.

Theorem 3. *Let Assumption 1 be satisfied. We then have*

$$\sqrt{n}\left(\widehat{\beta} - \beta^0\right) = \overline{W}_\infty^{-1}\, U + o_P(1),$$

where the d_β-vector U has elements

$$\begin{aligned}
U_k := \frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} \Bigg\{ & \ell_{ij}^{*\,\beta_k} - \mathbb{E}\left[\left(\ell_{ij}^{*\,\beta_k\alpha_i}\right)' \left(\sum_{h\in\mathcal{D}_i} \overline{\ell}_{ih}^{\alpha_i\alpha_i} \right)^{-1} \ell_{ij}^{\alpha_i} \right] - \mathbb{E}\left[\left(\ell_{ij}^{*\,\beta_k\gamma_j}\right)' \left(\sum_{h\in\mathcal{D}_j} \overline{\ell}_{hj}^{\gamma_j\gamma_j} \right)^{-1} \ell_{ij}^{\gamma_j} \right] \\
& + \frac{1}{2}\, \mathbb{E}\left[\left(\ell_{ij}^{\alpha_i}\right)' \left(\sum_{h\in\mathcal{D}_i} \overline{\ell}_{ih}^{\alpha_i\alpha_i} \right)^{-1} \left(\sum_{h\in\mathcal{D}_i} \overline{\ell}_{ih}^{*\,\beta_k\alpha_i\alpha_i} \right) \left(\sum_{h\in\mathcal{D}_i} \overline{\ell}_{ih}^{\alpha_i\alpha_i} \right)^{-1} \ell_{ij}^{\alpha_i} \right] \\
& + \frac{1}{2}\, \mathbb{E}\left[\left(\ell_{ij}^{\gamma_j}\right)' \left(\sum_{h\in\mathcal{D}_j} \overline{\ell}_{hj}^{\gamma_j\gamma_j} \right)^{-1} \left(\sum_{h\in\mathcal{D}_j} \overline{\ell}_{hj}^{*\,\beta_k\gamma_j\gamma_j} \right) \left(\sum_{h\in\mathcal{D}_j} \overline{\ell}_{hj}^{\gamma_j\gamma_j} \right)^{-1} \ell_{ij}^{\gamma_j} \right] \Bigg\}.
\end{aligned}$$

Furthermore, if also Assumption 2 holds, then

$$\begin{aligned}
\widehat{\delta}-\delta^0 = \left(\overline{\Delta}^{*\beta}\right)'(\widehat{\beta}-\beta^0) + \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \Bigg\{ & \psi_i^{(\alpha)\prime}\ell_{ij}^{*\alpha_i} + \psi_j^{(\gamma)\prime}\ell_{ij}^{*\gamma_j} \\
& - \mathbb{E}\left[\left(\Delta_{ij}^{\alpha_i} + \ell_{ij}^{\alpha_i\alpha_i}\psi_i^{(\alpha)} + \ell_{ij}^{\alpha_i'\gamma_j}\psi_j^{(\gamma)}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right] \\
& - \mathbb{E}\left[\left(\Delta_{ij}^{\gamma_j} + \ell_{ij}^{\gamma_j\alpha_i}\psi_i^{(\alpha)} + \ell_{ij}^{\gamma_j\gamma_j}\psi_j^{(\gamma)}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right] \\
& + \frac{1}{2}\,\mathbb{E}\left[\left(\ell_{ij}^{\alpha_i}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\left(\sum_{h\in\mathcal{D}_i}\overline{\Delta}_{ih}^{\#\,\alpha_i\alpha_i}\right)\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right] \\
& + \frac{1}{2}\,\mathbb{E}\left[\left(\ell_{ij}^{\gamma_j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\left(\sum_{h\in\mathcal{D}_j}\overline{\Delta}_{hj}^{\#\,\gamma_j\gamma_j}\right)\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right]\Bigg\} + o_P\left(1/\sqrt{n}\right),
\end{aligned}$$

where the $d_\alpha \times d_\alpha$ *matrices* $\overline{\Delta}_{ij}^{\#\,\alpha_i\alpha_i}$ *and the* $d_\gamma \times d_\gamma$ *matrices* $\overline{\Delta}_{ij}^{\#\,\gamma_j\gamma_j}$ *are given by*

$$\begin{aligned}
\overline{\Delta}_{ij}^{\#\,\alpha_i\alpha_i} &= \overline{\Delta}_{ij}^{\alpha_i\alpha_i} + \sum_{g=1}^{d_\alpha}\overline{\ell}_{ij}^{\alpha_i\alpha_i\alpha_{ig}}\psi_{ig}^{(\alpha)} + \sum_{g=1}^{d_\gamma}\overline{\ell}_{ij}^{\alpha_i\alpha_i\gamma_{jg}}\psi_{jg}^{(\gamma)}, \\
\overline{\Delta}_{ij}^{\#\,\gamma_j\gamma_j} &= \overline{\Delta}_{ij}^{\gamma_j\gamma_j} + \sum_{g=1}^{d_\alpha}\overline{\ell}_{ij}^{\gamma_j\gamma_j\alpha_{ig}}\psi_{ig}^{(\alpha)} + \sum_{g=1}^{d_\gamma}\overline{\ell}_{ij}^{\gamma_j\gamma_j\gamma_{jg}}\psi_{jg}^{(\gamma)}.
\end{aligned}$$

Proof. # Expansion of $\widehat{\beta}$. Our assumptions together with results of Lemma 1, 2 and Lemma 3 guarantee that the conditions of Theorem B.1 and Corollary B.2 in Fernández-Val and Weidner (2016) are satisfied, so that by applying that corollary we have

$$\sqrt{n}(\widehat{\beta}-\beta^0) = \overline{W}_\infty^{-1}U + o_P(1),$$

where $U = U^{(0)} + U^{(1)}$, with

$$\begin{aligned}
U^{(0)} &= \mathcal{L}^\beta + \overline{\mathcal{L}}^{\beta\phi}\,\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi = \mathcal{L}^{*\beta} = \frac{1}{n^{1/2}}\sum_{(i,j)\in\mathcal{D}}\overline{\ell}_{ij}^{*\beta}, \\
U^{(1)} &= \widetilde{\mathcal{L}}^{\beta\phi}\,\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi - \overline{\mathcal{L}}^{\beta\phi}\,\overline{\mathcal{H}}^{-1}\,\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\mathcal{L}^\phi + \frac{1}{2}\sum_{g=1}^{d_\phi}\left(\overline{\mathcal{L}}^{\beta\phi\phi_g} + \overline{\mathcal{L}}^{\beta\phi}\,\overline{\mathcal{H}}^{-1}\overline{\mathcal{L}}^{\phi\phi\phi_g}\right)[\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi]_g\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi \\
&= \widetilde{\mathcal{L}}^{*\beta\phi}\,\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi + \frac{1}{2}\sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{*\beta\phi\phi_g}\,[\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi]_g\,\overline{\mathcal{H}}^{-1}\mathcal{L}^\phi.
\end{aligned}$$

Here, tilde symbols indicate deviations from expectation, for example, $\widetilde{\mathcal{L}}^{\beta\phi} = \mathcal{L}^{\beta\phi} - \overline{\mathcal{L}}^{\beta\phi}$, with $\overline{\mathcal{L}}^{\beta\phi} = \mathbb{E}\mathcal{L}^{\beta\phi}$. Analogous to the proof of Theorem C.1 in Fernández-Val and Weidner (2016), and also using the above Lemma 2 again, one can then show that the terms in $U^{(1)}$ only contribute asymptotic bias, namely

$$
\begin{aligned}
\widetilde{\mathcal{L}}^{*\,\beta\phi}\,\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi} &= \mathbb{E}\left[\widetilde{\mathcal{L}}^{*\,\beta\phi}\,\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi}\right] + o_P(1) \\
&= \mathbb{E}\left[\widetilde{\mathcal{L}}^{*\,\beta\alpha}\,\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right] + \mathbb{E}\left[\widetilde{\mathcal{L}}^{*\,\beta\gamma}\,\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right] + o_P(1), \\
\frac{1}{2}\sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{*\,\beta\phi\phi_g}\,[\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi}]_g\,\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi} &= \mathbb{E}\left[\frac{1}{2}\sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{*\,\beta\phi\phi_g}\,[\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi}]_g\,\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi}\right] + o_P(1) \\
&= \mathbb{E}\left[\frac{1}{2}\sum_{g=1}^{Id_\alpha}\overline{\mathcal{L}}^{*\,\beta\alpha\alpha_g}\,\left[\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right]_g\,\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right] \\
&\quad + \mathbb{E}\left[\frac{1}{2}\sum_{g=1}^{Jd_\gamma}\overline{\mathcal{L}}^{*\,\beta\gamma\gamma_g}\,\left[\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right]_g\,\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right] + o_P(1).
\end{aligned}
$$

In component notation we can now rewrite the above terms as follows (remember that we define the Hessian matrix $\overline{\mathcal{H}}$ with a negative sign)

$$
\begin{aligned}
\mathcal{L}^{\beta} &= \frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\ell_{ij}^{*\,\beta_k} \\
\mathbb{E}\left[\widetilde{\mathcal{L}}^{*\,\beta\alpha}\,\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right] &= -\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\mathbb{E}\left[\left(\ell_{ij}^{*\,\beta_k\alpha_i}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right], \\
\mathbb{E}\left[\widetilde{\mathcal{L}}^{*\,\beta\gamma}\,\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right] &= -\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\mathbb{E}\left[\left(\ell_{ij}^{*\,\beta_k\gamma_j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right],
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}\sum_{g=1}^{Id_\alpha}\overline{\mathcal{L}}^{*\,\beta\alpha\alpha_g}\,\left[\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right]_g\,\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right] \\
&\qquad = \frac{1}{2}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\mathbb{E}\left[\left(\ell_{ij}^{\alpha_i}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{*\,\beta_k\alpha_i\alpha_i}\right)\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right] \\
&\mathbb{E}\left[\frac{1}{2}\sum_{g=1}^{Jd_\gamma}\overline{\mathcal{L}}^{*\,\beta\gamma\gamma_g}\,\left[\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right]_g\,\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right] \\
&\qquad = \frac{1}{2}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\mathbb{E}\left[\left(\ell_{ij}^{\gamma_j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{*\,\beta_k\gamma_j\gamma_j}\right)\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right].
\end{aligned}
$$

Combining the above gives the expansion for $\widehat{\beta}-\beta^0$ in the theorem.

Expansion of $\widehat{\delta}$. Again, our assumptions and lemmas guarantee that the conditions of Theorem B.4 in Fernández-Val and Weidner (2016) are satisfies, so that by applying that theorem we have

$$\begin{aligned}
\widehat{\delta} - \delta &= \left(\overline{\Delta}^{\beta} + \overline{\mathcal{L}}^{\beta\phi}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)'(\widehat{\beta} - \beta^0) + U_{\Delta}^{(0)} + U_{\Delta}^{(1)} + o_P\left(1/\sqrt{n}\right) \\
&= \left(\overline{\Delta}^{*\,\beta}\right)'(\widehat{\beta} - \beta^0) + U_{\Delta}^{(0)} + U_{\Delta}^{(1)} + o_P\left(1/\sqrt{n}\right),
\end{aligned}$$

where

$$\begin{aligned}
U_{\Delta}^{(0)} &= \mathcal{L}^{\phi\,\prime}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}, \\
U_{\Delta}^{(1)} &= \mathcal{L}^{\phi\,\prime}\,\overline{\mathcal{H}}^{-1}\,\widetilde{\Delta}^{\phi} - \mathcal{L}^{\phi\,\prime}\,\overline{\mathcal{H}}^{-1}\,\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi} + \tfrac{1}{2}\,\mathcal{L}^{\phi\,\prime}\overline{\mathcal{H}}^{-1}\left[\overline{\Delta}^{\phi\phi} + \sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\phi\phi\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_g\right]\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi}.
\end{aligned}$$

Again, following the logic in the proof of Theorem C.1 in Fernández-Val and Weidner (2016) one finds that $U_{\Delta}^{(1)}$ only contributes asymptotic bias, namely

$$\begin{aligned}
\mathcal{L}^{\phi\,\prime}\,\overline{\mathcal{H}}^{-1}\,\widetilde{\Delta}^{\phi} - \mathcal{L}^{\phi\,\prime}\,\overline{\mathcal{H}}^{-1}\,\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi} &= \mathbb{E}\left[\mathcal{L}^{\phi\,\prime}\,\overline{\mathcal{H}}^{-1}\left(\widetilde{\Delta}^{\phi} - \widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)\right] + o_P\left(1/\sqrt{n}\right) \\
&= \mathbb{E}\left\{\mathcal{L}^{\alpha\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\left[\widetilde{\Delta}^{\alpha} - \left(\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_{(\alpha)}\right]\right\} \\
&\quad + \mathbb{E}\left\{\mathcal{L}^{\gamma\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\left[\widetilde{\Delta}^{\gamma} - \left(\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_{(\gamma)}\right]\right\} + o_P\left(1/\sqrt{n}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\tfrac{1}{2}\,\mathcal{L}^{\phi\,\prime}\overline{\mathcal{H}}^{-1}\left[\overline{\Delta}^{\phi\phi} + \sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\phi\phi\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_g\right]\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi} \\
&= \mathbb{E}\left\{\tfrac{1}{2}\,\mathcal{L}^{\phi\,\prime}\overline{\mathcal{H}}^{-1}\left[\overline{\Delta}^{\phi\phi} + \sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\phi\phi\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_g\right]\overline{\mathcal{H}}^{-1}\mathcal{L}^{\phi}\right\} + o_P\left(1/\sqrt{n}\right) \\
&= \mathbb{E}\left\{\tfrac{1}{2}\,\mathcal{L}^{\alpha\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\left[\overline{\Delta}^{\alpha\alpha} + \sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\alpha\alpha\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_g\right]\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right\} \\
&\quad + \mathbb{E}\left\{\tfrac{1}{2}\,\mathcal{L}^{\gamma\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\left[\overline{\Delta}^{\gamma\gamma} + \sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\gamma\gamma\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_g\right]\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right\} + o_P\left(1/\sqrt{n}\right).
\end{aligned}$$

In component notation we can now rewrite the above terms as follows (again, remember that we define the Hessian matrix $\overline{\mathcal{H}}$ with a negative sign)

$$\mathbb{E}\left\{\mathcal{L}^{\alpha\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\left[\widetilde{\Delta}^{\alpha}-\left(\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_{(\alpha)}\right]\right\}$$
$$=-\mathbb{E}\left[\left(\Delta^{\alpha_i}_{ij}+\ell^{\alpha_i\alpha_i}_{ij}\psi^{(\alpha)}_{i}+\ell^{\alpha_i'\gamma_j}_{ij}\psi^{(\gamma)}_{j}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\ell^{\alpha_i}_{ij}\right],$$
$$\mathbb{E}\left\{\mathcal{L}^{\gamma\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\left[\Delta^{\gamma}-\left(\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_{(\gamma)}\right]\right\}$$
$$=-\mathbb{E}\left[\left(\Delta^{\gamma_j}_{ij}+\ell^{\gamma_j\alpha_i}_{ij}\psi^{(\alpha)}_{i}+\ell^{\gamma_j\gamma_j}_{ij}\psi^{(\gamma)}_{j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}^{\gamma_j\gamma_j}_{hj}\right)^{-1}\ell^{\gamma_j}_{ij}\right],$$

$$\mathbb{E}\left\{\tfrac{1}{2}\,\mathcal{L}^{\alpha\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\left[\overline{\Delta}^{\alpha\alpha}+\sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\alpha\alpha\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_{g}\right]\left(\overline{\mathcal{H}}^{*}_{(\alpha\alpha)}\right)^{-1}\mathcal{L}^{\alpha}\right\}$$
$$=\frac{1}{2}\,\mathbb{E}\left[\left(\ell^{\alpha_i}_{ij}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\left(\sum_{h\in\mathcal{D}_i}\overline{\Delta}^{\#\,\alpha_i\alpha_i}_{ih}\right)\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\ell^{\alpha_i}_{ij}\right],$$
$$\mathbb{E}\left\{\tfrac{1}{2}\,\mathcal{L}^{\gamma\,\prime}\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\left[\overline{\Delta}^{\gamma\gamma}+\sum_{g=1}^{d_\phi}\overline{\mathcal{L}}^{\gamma\gamma\phi_g}\left(\overline{\mathcal{H}}^{-1}\,\overline{\Delta}^{\phi}\right)_{g}\right]\left(\overline{\mathcal{H}}^{*}_{(\gamma\gamma)}\right)^{-1}\mathcal{L}^{\gamma}\right\}$$
$$=\frac{1}{2}\,\mathbb{E}\left[\left(\ell^{\gamma_j}_{ij}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}^{\gamma_j\gamma_j}_{hj}\right)^{-1}\left(\sum_{h\in\mathcal{D}_j}\overline{\Delta}^{\#\,\gamma_j\gamma_j}_{hj}\right)\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}^{\gamma_j\gamma_j}_{hj}\right)^{-1}\ell^{\gamma_j}_{ij}\right].$$

Combining the above gives the expansion for $\widehat{\delta}-\delta^0$ in the theorem. ■

A.6 Proof of Theorem 1 and 2

Proof of Theorem 1. According to Theorem 3 we have $\sqrt{n}\left(\widehat{\beta}-\beta^0\right)=\overline{W}_{\infty}^{-1}U+o_P(1)$. The first term in U is $\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\ell^{*\,\beta}_{ij}$, where in main text notation we have $\ell^{*\,\beta}_{ij}=\partial_z\ell_{ij}\tilde{X}_{ij}$. Assumption 1(i) guarantees that $\ell^{*\,\beta}_{ij}$ has mean zero (a linear combination of scores evaluated at the true parameters) and is either independent across all (i,j), or only correlated within pairs (i,j) and (j,i). This term therefore only contributes variance, no bias, to the limiting distribution of $\widehat{\beta}$. Applying the Lindeberg-Levy CLT and the Cramer-Wold device we find

$$\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\ell^{*\,\beta}_{ij}\to_d\mathcal{N}\left(0,\overline{\Sigma}_{\infty}\right),$$

where for the fully independent case (a) in Assumption 1(i) we have[8]

$$\overline{\Sigma}_{\infty} = \operatorname*{plim}_{I,J\to\infty} \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left(\ell^{*\beta}_{ij}\right)\left(\ell^{*\beta}_{ij}\right)' = \operatorname*{plim}_{I,J\to\infty} \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left(-\ell^{*\beta\beta}_{ij}\right) = \overline{W}_{\infty}.$$

Thus, in this case (a) the asymptotic variance of $\widehat{\beta}$ simplifies to $W_{\infty}^{-1}\overline{\Sigma}_{\infty}\overline{W}_{\infty}^{-1} = \overline{W}_{\infty}^{-1}$. For case (b) of Assumption 1(i) we have

$$\begin{aligned}
\overline{\Sigma}_{\infty} &= \operatorname*{plim}_{I,J\to\infty} \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \left[\mathbb{E}\left(\ell^{*\beta}_{ij}\right)\left(\ell^{*\beta}_{ij}\right)' + \mathbb{E}\left(\ell^{*\beta}_{ij}\right)\left(\ell^{*\beta}_{ji}\right)'\right] \\
&= \operatorname*{plim}_{I,J\to\infty} \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left\{\left(\partial_z \ell_{ij}\tilde{X}_{ij} + \partial_z \ell_{ji}\tilde{X}_{ji}\right)\partial_z \ell_{ij}\tilde{X}'_{ij}\right\},
\end{aligned}$$

where we used $\ell^{*\beta}_{ij} = \partial_z \ell_{ij}\tilde{X}_{ij}$. This is the formula for $\overline{\Sigma}_{\infty}$ given in Theorem 3, and this formula covers both case (a) and case (b), because independence across pairs $(i,j) \leftrightarrow (j,i)$ is of course a special case of dependence across those pairs.

All the remaining terms in U contribute asymptotic bias but no variance. We consider case (a) of Assumption 1(i) in the following, but one can easily verify that the additional bias terms stemming from correlation across pairs $(i,j) \leftrightarrow (j,i)$ are asymptotically negligible, so that the same asymptotic bias expressions are obtained in case (b).

Using $\ell^{*\beta_k\alpha_i}_{ij} = \gamma^0_j \partial_{z^2}\ell_{ij}\tilde{X}_{ij,k}$ and $\overline{\ell}^{\alpha_i\alpha_i}_{ih} = \gamma^0_j\gamma^{0\prime}_j \partial_{z^2}\ell_{ij}$ and $\ell^{\alpha_i}_{ij} = \gamma^0_j \partial_z \ell_{ij}$ we obtain

$$\mathbb{E}\left[\left(\ell^{*\beta_k\alpha_i}_{ij}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\ell^{\alpha_i}_{ij}\right] = \gamma^{0\prime}_j\left(\sum_{h\in\mathcal{D}_i}\gamma^0_h\gamma^{0\prime}_h\partial_{z^2}\ell_{ih}\right)^{-1}\gamma^0_j\,\mathbb{E}\left(\partial_z\ell_{ij}\partial_{z^2}\ell_{ij}\right),$$

and also using $\overline{\ell}^{*\beta_k\alpha_i\alpha_i}_{ih} = \gamma^0_j\gamma^{0\prime}_j\partial_{z^3}\ell_{ij}\tilde{X}_{ij,k}$ and the Bartlett identity $\mathbb{E}\ell^{\alpha_i}_{ij}\left(\ell^{\alpha_i}_{ij}\right)' = -\overline{\ell}^{\alpha_i\alpha_i}_{ij}$,

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{D}} \mathbb{E}\left[\left(\ell^{\alpha_i}_{ij}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{*\beta_k\alpha_i\alpha_i}_{ih}\right)\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\ell^{\alpha_i}_{ij}\right] \\
&= -\sum_{i=1}^{I}\operatorname{Tr}\left[\left(\sum_{j\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ij}\right)^{-1}\left(\sum_{j\in\mathcal{D}_i}\overline{\ell}^{*\beta_k\alpha_i\alpha_i}_{ij}\right)\right] = -\sum_{(i,j)\in\mathcal{D}}\operatorname{Tr}\left[\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}^{\alpha_i\alpha_i}_{ih}\right)^{-1}\overline{\ell}^{*\beta_k\alpha_i\alpha_i}_{ij}\right] \\
&= -\sum_{(i,j)\in\mathcal{D}}\gamma^{0\prime}_j\left(\sum_{h\in\mathcal{D}_i}\gamma^0_h\gamma^{0\prime}_h\partial_{z^2}\ell_{ih}\right)^{-1}\gamma^0_j\,\mathbb{E}\left(\partial_{z^3}\ell_{ij}\tilde{X}_{ij}\right),
\end{aligned}$$

[8] Here, we also used the Bartlett identity $\mathbb{E}\left(\ell^{*\beta}_{ij}\right)\left(\ell^{*\beta}_{ij}\right)' = \mathbb{E}\left(-\ell^{*\beta\beta}_{ij}\right)$.

and therefore

$$\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\left\{-\mathbb{E}\left[\left(\ell_{ij}^{*\beta_k\alpha_i}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right]\right.$$
$$\left.+\frac{1}{2}\,\mathbb{E}\left[\left(\ell_{ij}^{\alpha_i}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{*\beta_k\alpha_i\alpha_i}\right)\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right]\right\}$$
$$=-\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\gamma_j^{0\prime}\left(\sum_{h\in\mathcal{D}_i}\gamma_h^0 j\gamma_h^{0\prime}\partial_{z^2}\ell_{ih}\right)^{-1}\gamma_j^0\,\mathbb{E}\left(\partial_z\ell_{ij}\partial_{z^2}\ell_{ij}+\frac{1}{2}\partial_{z^3}\ell_{ij}\tilde{X}_{ij}\right)$$
$$=\sqrt{n}\,\frac{I}{n}\underbrace{\left[-\frac{1}{I}\sum_{i=1}^{I}\frac{1}{|\mathcal{D}_i|}\sum_{j\in\mathcal{D}_i}\gamma_j^{0\prime}\left(\frac{1}{|\mathcal{D}_i|}\sum_{h\in\mathcal{D}_i}\gamma_h^0\gamma_h^{0\prime}\partial_{z^2}\ell_{ih}\right)^{-1}\gamma_j^0\,\mathbb{E}\left(\partial_z\ell_{ij}\partial_{z^2}\ell_{ij}+\frac{1}{2}\partial_{z^3}\ell_{ij}\tilde{X}_{ij}\right)\right]}_{\to_P\overline{B}_\infty}.$$

Analogously we obtain

$$\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\left\{-\mathbb{E}\left[\left(\ell_{ij}^{*\beta_k\gamma_j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right]\right.$$
$$\left.+\frac{1}{2}\,\mathbb{E}\left[\left(\ell_{ij}^{\gamma_j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{*\beta_k\gamma_j\gamma_j}\right)\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right]\right\}$$
$$=\sqrt{n}\,\frac{J}{n}\underbrace{\left[-\frac{1}{J}\sum_{j=1}^{J}\frac{1}{|\mathcal{D}_j|}\sum_{i\in\mathcal{D}_j}\alpha_i^{0\prime}\left(\frac{1}{|\mathcal{D}_j|}\sum_{h\in\mathcal{D}_j}\alpha_h^0\alpha_h^{0\prime}\partial_{z^2}\ell_{hj}\right)^{-1}\alpha_i^0\,\mathbb{E}\left(\partial_z\ell_{ij}\partial_{z^2}\ell_{ij}+\frac{1}{2}\partial_{z^3}\ell_{ij}\tilde{X}_{ij}\right)\right]}_{\to_P\overline{D}_\infty}$$

Combining the above gives the statement of the theorem. ■

Proof of Theorem 2. Analogous to the proof of Theorem 1 we need to translate the stochastic expansion of $\widehat{\delta}$ in Theorem 3 into the notation used in the main text. We have $\left(\overline{\Delta}^{*\beta}\right)' \to_P \overline{(D_\beta\Delta)}_\infty$ and $\Psi_{ij}=-\psi_i^{(\alpha)\prime}\gamma_j^0-\psi_j^{(\gamma)\prime}\alpha_i^0$, and therefore find for the variance terms that

$$\underbrace{\left(\overline{\Delta}^{*\beta}\right)'\overline{W}_\infty^{-1}\ell_{ij}^{*\beta}}_{=\overline{(D_\beta\Delta)}_\infty\overline{W}_\infty^{-1}\partial_z\ell_{ij}\tilde{X}_{ij}}+\underbrace{\psi_i^{(\alpha)\prime}\ell_{ij}^{*\alpha_i}+\psi_j^{(\gamma)\prime}\ell_{ij}^{*\gamma_j}}_{=-\Psi_{ij}\partial_z\ell_{ij}}=\Gamma_{ij}.$$

Analogous to the proof of Theorem 1 one can show for the bias terms that

$$\frac{1}{I}\sum_{(i,j)\in\mathcal{D}}\left\{-\mathbb{E}\left[\left(\Delta_{ij}^{\alpha_i}+\ell_{ij}^{\alpha_i\alpha_i}\psi_i^{(\alpha)}+\ell_{ij}^{\alpha_i'\gamma_j}\psi_j^{(\gamma)}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right]\right.$$
$$\left.+\frac{1}{2}\,\mathbb{E}\left[\left(\ell_{ij}^{\alpha_i}\right)'\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\left(\sum_{h\in\mathcal{D}_i}\overline{\Delta}_{ih}^{\#\,\alpha_i\alpha_i}\right)\left(\sum_{h\in\mathcal{D}_i}\overline{\ell}_{ih}^{\alpha_i\alpha_i}\right)^{-1}\ell_{ij}^{\alpha_i}\right]\right\}\to_P \overline{B}_{\infty}^{\delta},$$

and

$$\frac{1}{J}\sum_{(i,j)\in\mathcal{D}}\left\{-\mathbb{E}\left[\left(\Delta_{ij}^{\gamma_j}+\ell_{ij}^{\gamma_j\alpha_i}\psi_i^{(\alpha)}+\ell_{ij}^{\gamma_j\gamma_j}\psi_j^{(\gamma)}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right]\right.$$
$$\left.+\frac{1}{2}\,\mathbb{E}\left[\left(\ell_{ij}^{\gamma_j}\right)'\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\left(\sum_{h\in\mathcal{D}_j}\overline{\Delta}_{hj}^{\#\,\gamma_j\gamma_j}\right)\left(\sum_{h\in\mathcal{D}_j}\overline{\ell}_{hj}^{\gamma_j\gamma_j}\right)^{-1}\ell_{ij}^{\gamma_j}\right]\right\}\to_P \overline{D}_{\infty}^{\delta}.$$

Using the above and the expansion in Theorem 3 gives the statement of Theorem 2. ■