R&D spillover effects on firm innovation - A spatial approach *

Emmanuel S. Tsyawo†

Latest version here

September 2019

Abstract

Quantifying R&D spillover effects on innovation requires a spatial matrix that characterises the strength of connectivity between firms. Estimates can be biased if the spatial matrix is misspecified. This paper proposes a parsimonious approach to estimating the spatial matrix, when the spatial matrix is partly or fully unknown, alongside parameters and quantifies R&D spillovers on innovation. The approach is applicable to linear and non-linear models and allows asymmetry and time-variation in the spatial matrix. In an application of the approach to the negative binomial model, this paper establishes consistency and asymptotic normality of the MLE under conditional independence and conditional strong-mixing assumptions on the outcome variable. On firm innovation, we find positive spillover and private effects of R&D. We provide evidence of time-variation and asymmetry in the interaction structure between firms and find that geographic proximity and product market proximity are relevant. Moreover, the strength of connectivity between firms is not limited to often-assumed notions of closeness; it is also linked to past R&D and patenting behaviour of firms.

Keywords: R&D spillovers, innovation, spatial dependence, spatial matrix

JEL classification: C21, L25

*Our gratitude goes to Brantly Callaway, Oleg Rytchkov, Charles E. Swanson, Austin Bean, Catherine Maclean, Pedro Silos, and all participants of the Temple University Econ Summer Job Market Boot-camp for helpful comments. This project is supported by the Temple University Economics Department Summer (2019) Research Grant.

†Department of Economics, Temple University. Email: estsyawo@temple.edu

1 Introduction

Quantifying the spillover effects of research and development (R&D) on firm innovation, productivity, costs, and growth is a major pursuit in the empirical industrial organisation literature. Firms' research and development (R&D) efforts are known to have spillover effects on other firms' outcomes (e.g., Grillitsch and Nilsson (2015), Bloom, Schankerman, and Van Reenen (2013), and Cincera (1997)). Firms engage in R&D not only to create knowledge but also to boost capacity in order to use knowledge created by other firms (Cohen and Levinthal, 1989). While there appears to be a consensus on the existence and importance of R&D spillovers, existing approaches to quantifying R&D spillover effects remain varied. Knowledge spillovers have policy implications and accurately quantifying them aids to fine-tune policy,[1] promote innovation, and accelerate technological progress.

A crucial component to quantifying spillover effects is the spillover measure. The spillover measure corresponding to a firm is obtained by weighting the R&D of other firms using a spatial matrix that defines the strength of connectivity between firms. In the literature, the spatial matrix is often assumed known (e.g., Bloom, Schankerman, and Van Reenen (2013)). We note that misspecifying the spatial matrix (and spillover measure by extension) leads to misleading conclusions and policy recommendations (e.g., Bhattacharjee and Jensen-Butler (2013)). In light of the foregoing challenge, this paper introduces a flexible approach that models and estimates the spatial matrix from panel data in order to quantify R&D spillover effects on innovation.

The method we propose models the structure of interactions between firms across time using spatial covariates and a parameter vector of finite length. The framework is applicable to a wide class of linear and non-linear models. This is particularly crucial for quantifying spillovers and spatial dependence when the appropriate model is not linear. Our approach

[1]Spillover effects, e.g., positive externalities imply under-investment in R&D from a social perspective and a need to subsidise firm R&D effort via policies such as R&D tax credits. In assessing knowledge spillovers of firms, Bloom, Schankerman, and Van Reenen (2013) and Hoshino (2017) finds limited knowledge spillovers of small firms and recommends a reconsideration of higher R&D tax credits given to small firms.

is parsimonious, data-driven, and testable. It allows a time-varying and asymmetric spatial matrix that admits several spillover covariates simultaneously. The relevance of spatial covariates in the spatial matrix is statistically assessable and obviates the arbitrary choice of metric, spatial matrix, or both. We are able to identify spillover covariates and conduct hypothesis tests on their significance in order to establish their relevance. Identification of spillover sources (measured by spillover covariates, e.g., geographic distance) is particularly suitable for studies quantifying spillover effects as it sheds light on the channels of transmission of spillovers between firms. Our method is easy to implement and standard tools of M-estimation and inference are applicable. We contribute to the literature in two ways. First, we quantify spillover effects when the interaction structure is partly or fully unknown. Second, the approach generalises to a class of linear and non-linear models that admit linear predictor functions.

The paper applies the approach to the negative binomial model and provides results on the consistency and asymptotic normality of the maximum likelihood estimator. Under conditional stationarity (across time), conditional independence (across firms), and strong-mixing assumptions (over time) assumptions on the outcome, the consistency and asymptotic normality of the maximum likelihood estimator are proven. Because the econometric framework only requires conditional independence, the outcome across firms is allowed to be unconditionally dependent, and the outcome of firms across time are allowed to be conditionally strong-mixing.

Our empirical application studies spillover effects of R&D on firm innovation. In following the literature (e.g., Aghion, Van Reenen, and Zingales (2013) and Bloom, Schankerman, and Van Reenen (2013)), we use variables including citation-weighted patent counts (as outcome variable for innovation) and R&D stock (for R&D) in estimation. Due to the discreteness of citation-weighted patent counts, the approach is applied to the negative binomial model for estimation and inference. Our empirical results confirm the sensitivity of spillover measures to the choice of spatial covariate (and spatial matrix, by extension). We confirm the presence of positive spillover effects of R&D on firm innovation. Though the private effect of R&D on innovation is positive and statistically significant, it is dominated by the spillover effect. Besides geographic proximity and product market proximity, the other relevant spillover covariates are associated with past R&D and patenting behaviour of firms. The results on relevant spatial covariates confirm asymmetry and time-variation in the strength of connectivity between firms with respect to innovation. Also, the results confirm that the strength of connectivity between firms is not only tied to commonly used notions of proximity viz. geographic and product market.

The rest of the paper is organised as follows. Section 2 provides a review of the literature and section 3 describes the model and presents an identification result. Section 4 provides empirical and econometric extensions to the framework presented in this paper. Section 5 covers estimation algorithms and section 6 presents asymptotic results. The empirical study is conducted in section 7, and section 8 concludes. All proofs are relegated to appendices A and B.

2 Related Literature

Works in the literature construct R&D spillover measures in a number of ways; we identify five of them. A simple approach pools the R&D or knowledge stock of other firms in the same industry (e.g., Bernstein and Nadiri (1989) and Cincera (1997)).[2] A second approach employs industry distance to the frontier[3] or a measure of potential of knowledge transfer as a proxy for technological spillovers (e.g., Grillitsch and Nilsson (2015), Acemoglu et al. (2007), and Griffith, Redding, and Reenen (2004)). A third approach constructs spillover measures by using exports, imports, and foreign direct investment to weight domestic and foreign R&D stocks (e.g., Coe, Helpman, and Hoffmaister (2009)). Bloom, Schankerman,

[2] Cincera (1997) pools R&D investment at the manufacturing sector level and only allows intra-sector interactions.

[3] Griffith, Redding, and Reenen (2004) defines the frontier as the economy with the largest total factor productivity (TFP) in a given sector.

and Van Reenen (2013) uses Jaffe (1986)'s un-centred correlation coefficient approach to construct exogenous spatial matrices based on technology, product market, and geographic proximity. Last, a recent approach directly estimates the spatial matrix from data (e.g., Manresa (2013) and Soale and Tsyawo (2019)).

At the econometric level, much progress is realised in the last two decades with respect to modelling spillover effects and spatial dependence in general. A first category spatially lags the outcome variable, namely, the Spatial Autoregressive (SAR) model (Anselin, Le Gallo, and Jayet, 2008). A second category estimates the spatial matrix as a set of parameters. The second category requires high-dimensional estimation techniques whenever the number of time periods does not substantially exceed the number of firms (e.g., Manresa, 2013; Soale and Tsyawo, 2019; Lam and Souza, 2019). A third models individual elements of the spatial matrix up to a finite number of parameters and jointly estimates all parameters using panel data (Kapetanios, Mitchell, and Shin, 2014; Pinkse, Slade, and Brett, 2002). Our approach falls in this last category.

As regards constructing spillover measures and quantifying R&D spillovers, we note a number of shortcomings of existing approaches. Exogenously constructing spillover measures from pre-specified metrics can lead to misleading results if the choice of metric is not guided by theory. The underlying spatial matrix of the Jaffe (1986) measure is time-invariant and symmetric. Interaction between firms of different sizes and research capacities is generally not symmetric. In settings where the structure of interactions evolves through time, time-invariant spatial matrices are inadequate. In the presence of multiple sources of R&D spillovers, a researcher faces a *metric ambiguity* problem. Besides the often documented sources of knowledge transfers viz. geographic (Grillitsch and Nilsson, 2015; Lychagin, Pinkse, Slade, and Reenen, 2016), technology (Bloom, Schankerman, and Van Reenen, 2013), product market (Bloom, Schankerman, and Van Reenen, 2013), or input market proximity, knowledge spillovers can be explained by other forms of proximity that are often overlooked in the literature, e.g., social networks and labour mobility.[4]

Modelling spatial dependence is largely confined to linear models. Recent contributions to the literature on non-linear spatial models include Xu and Lee (2015) and Hoshino (2017). These approaches, however, suppose foreknowledge of the spatial matrix to other non-linear models. In this paper, innovation is measured using citation-weighted patent counts (see also Hall, Jaffe, and Trajtenberg (2005), Aghion, Van Reenen, and Zingales (2013), and Huang and Tsyawo (2018)). Besides citation-weighted patent counts, patent applications (e.g., Cincera (1997)), and a binary on collaboration in innovative activity (Grillitsch and Nilsson, 2015) are also used as measures of innovation. These measures are discrete and do require non-linear models viz. negative binomial, Poisson, and logit for estimation.

3 The model

Modelling spatial dependence using our approach spans a class of models that admit linear predictor functions, e.g., linear, quantile, logit, and probit regressions. In this paper, we focus on the negative binomial model as the outcome variable (Citation-weighted patent counts) is discrete. Let data $\{y_{it}, x_{it}, \mathbf{z}_{it}, \{\boldsymbol{d}_{ijt}\}_{j=1}^{N}, c_{it}\}_{i=1,t=1}^{N,T}$, be defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. The outcome y_{it}, covariates $\mathbf{z}_{it}$ and $\mathbf{x}_t = [x_{1t}, \ldots, x_{Nt}]'$ (where $\mathbf{x}_t$ generates spillovers, e.g., R&D), and spillover covariates $\boldsymbol{d}_{ijt}$ (e.g. geographic distance) are observed. c_{it} denotes unobserved heterogeneity. Define $\boldsymbol{D}_t \equiv \{\boldsymbol{d}_{ijt} : i = 1, \ldots, N,\ j \neq i\}$ and $\boldsymbol{D} \equiv \bigcup_{t=1}^{T} \boldsymbol{D}_t$. The conditional distribution assumption is stated in the following.

Assumption 1. *Conditional on $\mathcal{F}$, y_{it} follows the negative binomial distribution with a finite dispersion parameter $\eta^2 > 0$.*[5]

[4] Grillitsch and Nilsson (2015) finds that firms in peripheral regions with less access to local (geographical) knowledge spillovers compensate for the lack via collaborations with non-local firms. Bhattacharjee and Jensen-Butler (2013, p. 630) notes a number of non-significant weights between contiguous region pairs. These findings show that sources of knowledge spillovers are not always tied to often-assumed notions of closeness.

[5] For ease of exposure, η^2 is treated as unknown.

3.1 The conditional expectation

Under assumption 1, the conditional expectation and conditional variance, respectively are given by

$$E(y_{it}|\mathcal{F}) = E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) = \exp(\rho_0 + \rho_1 x_{it} + \rho_2 \sum_{j\neq i} x_{jt} w_{ijt}(\boldsymbol{\delta}) + \mathbf{z}_{it}\boldsymbol{\gamma}) \text{ and} \tag{3.1}$$

$$\operatorname{var}(y_{it}|\mathcal{F}) = \operatorname{var}(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) = E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) + \eta^2 E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) \tag{3.2}$$

where $w_{ijt}(\boldsymbol{\delta}) \equiv \dfrac{\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\sum_{j\neq i}^{N}\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}$, $\forall\ i = 1,\ldots,N,\ t = 1,\ldots,T$. The function $\exp(\cdot)$ can be replaced by any positive twice differentiable integrable function $f : \mathbb{R} \to \mathbb{R}_+$. The parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^{k_\theta}$, $k_\theta = k_\beta + k_\delta$ is partitioned into $\boldsymbol{B} \subset \mathbb{R}^{k_\beta}$ and $\boldsymbol{\Delta} \subset \mathbb{R}^{k_\delta}$ where $\boldsymbol{\beta} = [\rho_0, \rho_1, \rho_2, \boldsymbol{\gamma}']' \in \boldsymbol{B}$, $\boldsymbol{\delta} \in \boldsymbol{\Delta}$, $\boldsymbol{\Theta} \equiv \boldsymbol{B} \times \boldsymbol{\Delta}$, and $\boldsymbol{\theta} \equiv [\boldsymbol{\beta}', \boldsymbol{\delta}']'$. Define an $NT \times (k_\beta - 3)$ matrix $\mathbf{z}$ whose $((t-1)N+i)$'th row is $\mathbf{z}_{it}$ and an $NT \times 1$ vector $\mathbf{x}$ whose $((t-1)N+i)$'th element is x_{it}. Define an $NT \times NT$ block diagonal matrix $\boldsymbol{w}(\boldsymbol{\delta}) = diag[\boldsymbol{w}_1(\boldsymbol{\delta}), \ldots, \boldsymbol{w}_T(\boldsymbol{\delta})]$ where the (i,j)'th element of $\boldsymbol{w}_t(\boldsymbol{\delta})$, $t \in \{1,\ldots,T\}$ is $w_{ijt}(\boldsymbol{\delta})$ if $j \neq i$ and zero otherwise. The $NT \times k_\beta$ design matrix associated with the conditional expectation eq. (3.1) is given by $\boldsymbol{m} = [\mathbf{1}_{NT}, \mathbf{x}, \boldsymbol{w}\mathbf{x}, \mathbf{z}]$ where $\mathbf{1}_{NT}$ denotes an $NT \times 1$ vector of ones, and the dependence of $\boldsymbol{w}$ and $\boldsymbol{m}$ on $\boldsymbol{\delta}$ is suppressed for notational convenience.

Spillover covariates, denoted by the $k_\delta \times 1$ vector $\boldsymbol{d}_{ijt}$, are indicators of connectivity between firms and can contain several metrics (e.g. proximity in technological space and geographical proximity) and flexible functional forms of variables related to R&D, firm characteristics, or innovation but not necessarily tied to, for instance, geography or industry (e.g. lagged x_{it} and $x_{jt},\ j \neq i$ in, for example, Kapetanios, Mitchell, and Shin (2014)). $\boldsymbol{d}_{ijt} \in \boldsymbol{D}_t$ allows a flexible modelling of the weight function across observations and through time up to a parameter vector $\boldsymbol{\delta}$ of finite length k_δ. The formulation eq. (3.1) allows innovation of firm i to be impacted not only by own R&D x_{it} and characteristics $\mathbf{z}_{it}$, but also by the R&D $x_{jt},\ j \neq i$ of other firms.

The negative binomial is not the only model to which our approach is applicable; it applies to a class of models that admit linear predictor functions as in eq. (3.1). Examples include linear regression, quantile regression, logit, and Poisson models. To set the scope, consider a baseline functional of the distribution of outcome y_{it} conditional on observables

$$\nu(y_{it}|\mathcal{F}) = \nu(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) = g(\rho_0 + \rho_1 x_{it} + \rho_2 \sum_{j\neq i} x_{jt} w_{ijt}(\boldsymbol{\delta}) + \mathbf{z}_{it}\boldsymbol{\gamma}) \tag{3.3}$$

where $g : \mathbb{R} \to \mathbb{R}$ is known and differentiable, and the conditional functional can be the conditional expectation (e.g., the negative binomial model eq. (3.1)), the conditional quantile (quantile regression), or the conditional probability (distribution regression).

3.2 The objective function

The objective function has a sample average representation $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})$ and the parameter set of interest solves the following minimisation problem

$$[\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}']' = \underset{[\boldsymbol{\beta}', \boldsymbol{\delta}']' \in \boldsymbol{\Theta}}{\arg\min}\, \mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta}) \tag{3.4}$$

$q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})$ is the contribution of unit i at time t to the objective function $\mathcal{Q}_n(\cdot, \cdot)$ and $q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta}) = -\eta^{-2}\log[\eta^{-2}/(\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))] - y_{it}\log[\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})/(\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))]$ for a given finite $\eta^2 > 0$ where $\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) = \exp(\rho_0 + \rho_1 x_{it} + \rho_2 \sum_{j\neq i} x_{jt} w_{ijt}(\boldsymbol{\delta}) + \mathbf{z}_{it}\boldsymbol{\gamma})$ and $m_{it}(\boldsymbol{\delta})$ is the $((t-1)N+i)$'th row of $\boldsymbol{m}(\boldsymbol{\delta})$. $q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})$ of other models includes $q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta}) = (y_{it} - m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})^2$ for linear regression and $q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta}) = -(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})y_{it} - log(1 + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))$ for the logit model.

If $\boldsymbol{D}$ contains at least one time-varying element, $\boldsymbol{w}_t(\boldsymbol{\delta})$ for $\boldsymbol{\delta} \in \boldsymbol{\Delta}, \boldsymbol{\delta} \neq \mathbf{0}$ varies through time and the structure of spatial dependence need not be stable through time. Also, we do not assume $\boldsymbol{d}_{ijt} = \boldsymbol{d}_{ijt},\ j \neq i$ whence possible asymmetry in $\boldsymbol{w}(\boldsymbol{\delta})$. Properties of our approach like time-variation and asymmetry can be assessed statistically by testing the significance of elements in $\boldsymbol{\delta}$ corresponding to time-varying and asymmetric elements in $\boldsymbol{D}$.

3.3 Identification

We make the following assumptions.

Assumption 2. *(a)* $\boldsymbol{z} \equiv [\mathbf{1}_{NT}, \mathbf{x}, \mathbf{z}]$ *is full column rank almost surely (a.s.). (b)* $\boldsymbol{w}_t(\boldsymbol{\delta})$ *is non-singular a.s., for each* $t = 1, .., T$ *and for all* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$*. (c) The diagonal elements of* $\boldsymbol{w}_t(\boldsymbol{\delta})$ *are zeroes, for all* $t = 1, .., T$ *and* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$

Assumption 2(a) is a standard identification assumption. Assumption 2(b) only fails in pathological cases; it holds if there is enough variation in $\boldsymbol{D}_t$ for $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ and $t = 1, \ldots, T$. Note that $\mathbf{x} \neq \mathbf{0}$ a.s. follows from assumption 2(a), and it is necessary (coupled with assumption 2(b)) for $\boldsymbol{w}\mathbf{x}$ to possess independent variation from $\boldsymbol{z}$. These assumptions are verifiable.

The following assumptions are useful in bounding the design matrix $\boldsymbol{m}(\boldsymbol{\delta})$ for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$

Assumption 3. *(a) There exists a positive constant* κ_w*,* $0 < \kappa_w^8 < \infty$ *such that* $\sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \exp((\boldsymbol{d}_{ijt} - \bar{\boldsymbol{d}}_{i.t})'\boldsymbol{\delta}) \leq \kappa_w$ *a.s.,* $\bar{\boldsymbol{d}}_{i.t} \equiv N^{-1} \sum_{j \neq i} \boldsymbol{d}_{ijt}$ *for all* $\boldsymbol{d}_{ijt} \in \boldsymbol{D}_t$ *and* $t = 1, \ldots, T$*. (b) There exist positive constants* κ_y *and* κ_m*,* $1 \leq \kappa_y^8 < \infty$*, such that* $\sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} ||m_{it}(\boldsymbol{\delta})||_2 \leq \kappa_m$*,* $\sup_{\boldsymbol{\beta} \in \boldsymbol{B}} \kappa_m ||\boldsymbol{\beta}||_2 \leq 1/2 \log \kappa_y$ *a.s. for each* $i = 1, \ldots, N$*,* $t = 1, \ldots, T$*, and* $||\boldsymbol{\xi}||_p$ *denotes the p-norm,* $p \geq 1$ *applied to vectorised* $\boldsymbol{\xi}$*. (c) Each element in* $\boldsymbol{z} \equiv [\mathbf{1}_{NT}, \mathbf{x}, \mathbf{z}]$ *and* $\boldsymbol{D}$ *is bounded in absolute value, and there exist positive constants* κ_d*,* κ_z *such that* $\max_{\substack{j \in \{1,\ldots,N\} \\ t \in \{1,\ldots,T\} \\ j \neq i}} ||\boldsymbol{d}_{ijt}||_2 \leq \kappa_d$ *and* $||\boldsymbol{z}||_\infty \leq \kappa_z$ *a.s.,* $0 < \kappa_d^8 < \infty$ *and* $0 < \kappa_z^8 < \infty$*.*

Note that $\sup_{\boldsymbol{\beta} \in \boldsymbol{B}, \boldsymbol{\delta} \in \boldsymbol{\Delta}} |m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| \leq 1/2 \log \kappa_y$ a.s. follows from assumption 3(b) because $|m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| = ||m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}||_2 \leq ||m_{it}(\boldsymbol{\delta})||_2 ||\boldsymbol{\beta}||_2 \leq 1/2 \log \kappa_y$. By construction, the row sums of $\boldsymbol{w}(\boldsymbol{\delta})$ are bounded at 1. In the following lemma, we bound the column sums of $\boldsymbol{w}(\boldsymbol{\delta})$.

Lemma 1. *Under assumption 3(a),* $\sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \max_{\substack{j \in \{1,\ldots,N\} \\ t \in \{1,\ldots,T\}}} w_{ijt}(\boldsymbol{\delta}) \leq \frac{\kappa_w}{N}$ *a.s., for any* $i \neq j$*, and the column sums of* $\boldsymbol{w}(\boldsymbol{\delta})$ *are bounded in absolute value a.s. for all* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$*.*

In the following, we show the full column rank condition of the design matrix $\boldsymbol{m}$ holds for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ under assumption 2.

Lemma 2. *Under assumptions 2(a-c) and 3(b)* $\frac{1}{NT}\boldsymbol{m}(\boldsymbol{\delta})'\boldsymbol{m}(\boldsymbol{\delta})$ *is positive definite and bounded a.s. for all* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$.

Lemma 3. *Under assumption 3(b),* $E|\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})| < \infty$ *for all* $\boldsymbol{\beta} \in \boldsymbol{B}$ *and* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$.

Lemma 4 (Identification)**.** *Under the assumptions of lemma 2,* $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ *is identified.*

The uniqueness of the true parameter vector $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ can be established thanks to lemmata 1 to 4. The result is provided in the following theorem.

Theorem 1. *Under the assumptions of lemmata 1 to 4,* $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta}) \equiv E[\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})]$ *has a unique minimum at* $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$.

3.4 Parameters of interest

We explore parameters that are interesting from a policy perspective. Though these parameters can have varied interpretation by application (see section 4.1 for examples), we interpret them in the context of R&D spillover effects on innovation. In what follows, let x_{it} denote firm i's R&D in year t.

Private effect: The private effect of $R\&D$ measures the impact of a firm's R&D on its innovation. $\mathcal{PE}_i = T^{-1}\sum_{t=1}^{T}\frac{\partial E(y_{it}|\cdot)}{\partial x_{it}} = T^{-1}\sum_{t=1}^{T}\mathcal{PE}_{it} = \rho_1 T^{-1}\sum_{t=1}^{T}E(y_{it}|\cdot)$. The average private effect is given by $\mathcal{PE} = N^{-1}\sum_{i=1}^{N}\mathcal{PE}_i$. The baseline model (eq. (3.1)) does not allow heterogeneity in ρ_1 across firms unlike, Manresa (2013) and Soale and Tsyawo (2019, sect. 6). It does, however, allow variation through time and does not require a static interaction structure.

Spillover effect: The spillover effect measures the impact of firm j's R&D on i's innovation. The building block for computing spillover effects, taking account of pairwise interaction of firms $i, j, i \neq j$ and time t is $\mathcal{S}p\mathcal{E}_{ijt} = \frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \rho_2 w_{ijt}(\boldsymbol{\delta})E(y_{it}|\cdot)$. The spillover effect of a firm j's R&D is the effect it has on another firm i's, $j \neq i$, innovation. It is given by $\mathcal{S}p\mathcal{E}_{ij.} = T^{-1}\sum_{t=1}^{T}\mathcal{S}p\mathcal{E}_{ijt} = \rho_2 T^{-1}\sum_{t=1}^{T} w_{ijt}(\boldsymbol{\delta})E(y_{it}|\cdot)$. The spillover effect of other firms' R&D on firm i's innovation obtains as $\mathcal{S}p\mathcal{E}_{i..} = \sum_{j\neq i}\mathcal{S}p\mathcal{E}_{ij.}$, the spillover effect exerted by firm j on other firms' innovation is $\mathcal{S}p\mathcal{E}_{.j.} = \sum_{i\neq j}\mathcal{S}p\mathcal{E}_{ij.}$, and the average spillover effects across all firms is $\mathcal{S}p\mathcal{E} = N^{-1}\sum_{i=1}^{N}\mathcal{S}p\mathcal{E}_{i..} = N^{-1}\sum_{j=1}^{N}\mathcal{S}p\mathcal{E}_{.j.}$. A researcher interested in time-variation of average spillover effects can compute $\mathcal{S}p\mathcal{E}_{..t} = N^{-1}\sum_{i=1}^{N}\sum_{j\neq i}\mathcal{S}p\mathcal{E}_{ijt}$ for each $t \in \{1,\ldots,T\}$.

Social effect: The social effect of R&D on innovation sums private and spillover effects. The social effect generated by firm j's R&D at time t is $\mathcal{SE}_{.jt} = \sum_{i=1}^{N}\frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \frac{\partial E(y_{jt}|\cdot)}{\partial x_{jt}} + \sum_{i\neq j}\frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \mathcal{PE}_{jt} + \sum_{i\neq j}\mathcal{S}p\mathcal{E}_{ijt} = \mathcal{PE}_{jt} + \mathcal{S}p\mathcal{E}_{.jt}$. In a similar vein, the social effects of R&D (from all firms) received by firm i at time t is defined as $\mathcal{SE}_{i.t} = \sum_{j=1}^{N}\frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} + \sum_{j\neq i}\frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \mathcal{PE}_{it} + \sum_{j\neq i}\mathcal{S}p\mathcal{E}_{ijt} = \mathcal{PE}_{it} + \mathcal{S}p\mathcal{E}_{i.t}$. Average social effects across firms and time obtains as $\mathcal{SE} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{SE}_{i.t} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{SE}_{.jt}$.[6]

Strength of spatial dependence: ρ_2 controls the strength of the spatial dependence in our model; it can accommodate both weak and strong forms of spatial dependence[7] depending on the magnitude of ρ. A null hypothesis $H_o:\ \rho_2 = 0$ can be used to assess the strength of spatial dependence. This approach provides a way of testing the strength of spatial dependence in a specific model. Failing to reject H_o is indicative of a weak spatial dependence.

Relevant spatial covariates: By modelling the spatial matrix $\boldsymbol{w}$ as a function of observable spillover covariates in $\boldsymbol{D}$ up to a parameter vector $\boldsymbol{\delta}$ of finite length k_δ, our approach allows the evaluation of factors that drive spatial dependence i.e., test elements of $\boldsymbol{\delta}$ that

[6] Comparable approaches to computing structural parameters viz. private, spillover, and social effects are given in Bloom, Schankerman, and Van Reenen (2013, Sect. 6.5.1) and Manresa (2013, Sect. 5.5).

[7] This feature is also present in Kapetanios, Mitchell, and Shin (2014)'s model.

are statistically significant.[8] In a similar vein, the strength of time-variation or asymmetry in $\boldsymbol{w}$ is assessable by testing the significance (jointly or individually) of elements in $\boldsymbol{\delta}$ that correspond to time-varying or asymmetric elements in $\boldsymbol{D}$.

4 Extensions

4.1 Other possible empirical applications of the model

A notable aspect to our model, from the econometric perspective, is the ability to model spatial dependence in both linear and non-linear models when the spatial matrix is unknown. Also, it allows empirical determination of observable spillover covariates as well as an assessment of time-variation and asymmetry in the spatial matrix $\boldsymbol{w}$. Empirically, our method has applications beyond the study of R&D spillover effects on innovation. In the following, we illustrate, with empirical examples, other possible applications of our model.

R&D Spillovers on firm productivity: A major challenge with estimating R&D spillover effects on firm productivity is that spillovers may not be tied to any plausible metric, for example, geographic distance (Syverson, 2011). Like Lychagin, Pinkse, Slade, and Reenen (2016), our framework is applicable to study this problem by jointly accounting for all plausible sources of R&D spillovers (measured by the respective spatial covariates) in a production function framework. Unlike Bloom, Schankerman, and Van Reenen (2013), the spatial matrix is determined within the model and the interaction between firms over time can vary through time and be asymmetric. Our proposed framework can also be used to study R&D spillover effects on other interesting firm outcomes viz. market value and R&D factor demand.

[8] Lychagin, Pinkse, Slade, and Reenen (2016), for example, finds that R&D spillover effects on firm productivity via geographic and technological proximity (but not product market) are important.

Spillovers in demand behaviour: Spillovers in demand due to geographically varying prices, e.g., regional housing demand (Bhattacharjee and Jensen-Butler, 2013), rice demand (Case, 1991), and cigarette demand Kelejian and Piras, 2014 receive much attention in the literature. Markets are not perfectly segregated, and shocks to demand in a region impact demand behaviour in other regions. Our model can contribute to this literature by allowing time-variation, asymmetry, and other metrics (besides geographical contiguity) in determining the spatial matrix.

4.2 Extensions of the model

The approach we propose is not without drawbacks. Our approach requires observed metrics or spillover covariates, and it is unable to accommodate non-negative elements of the spatial matrix.[9] Also, the number of elements in the spatial matrix to be modelled grows in the order of $O(N^2T)$ which may require the dimension of $\boldsymbol{\delta}$ to grow. Because of the exponential function in $w_{ijt}(\boldsymbol{\delta})$, our approach may not accommodate sparse interaction structures well without further adjustment. Consistent estimation of the spatial matrix requires a balanced panel, and this can result in the loss of some firms. We propose extensions to the baseline model eq. (3.1) to deal with these challenges. Though we do not explore the computational and inferential challenges associated with the extensions, we present them.

Dynamic spatial dependence: Past outcomes of other firms may affect future outcomes. Such a phenomenon can be modelled by setting x_{it} to lagged outcome $y_{i,t-1}$.

$$E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) = \exp(\rho_0 + \rho_1 y_{i,t-1} + \rho_2 \sum_{j \neq i} y_{j,t-1} w_{ijt}(\boldsymbol{\delta}) + \mathbf{z}_{it}\boldsymbol{\gamma})$$

$E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t)$ above is a non-linear generalisation of Kapetanios, Mitchell, and Shin (2014)'s model. Like Cincera (1997, Sect. 4), dynamics in the spillover parameter can

[9] Soale and Tsyawo (2019), Pinkse, Slade, and Brett (2002), Bhattacharjee and Jensen-Butler (2013), and Manresa (2013) allows negative,zero, and positive weights.

be accommodated by including lagged x_{it}, x_{jt} and lagged spatial matrices in order to capture dynamics in, for example, the patenting process. In some applications, lagged outcome drives spatial dependence and can be used as a spatial covariate (e.g. Kapetanios, Mitchell, and Shin (2014)). More lags of the outcome can be included in the conditional expectation as covariates or spatial covariates provided there are sufficient degrees of freedom to estimate the model.

Multiple metrics: A particularly useful feature of our model lies in its ability to handle metric ambiguity by incorporating several metrics in $\boldsymbol{D}$. Spatial dependence in outcomes can emerge as an interplay of several metrics. For example, production-related spillovers from R&D can be simultaneously impacted by a number of factors linked to geographical, output market, input market, and technological distance. For example, Lychagin, Pinkse, Slade, and Reenen (2016) simultaneously assesses geographic, technological, and product market spaces as sources of R&D spillovers. In the case where there is uncertainty over the choice of relevant metrics, our formulation allows for a data-driven determination of the relevant subset of metrics and spatial covariates (or functions thereof) via hypothesis tests.

Prior information in spatial matrix: Where sparsity in the spatial matrix is justifiable, binaries on contiguity can be incorporated in the spatial matrix to induce sparsity. For instance, $w_{ijt}(\boldsymbol{\delta})$ can be weighted by $C(d_{ijt}) = 1$ if i and j are contiguous at time t and zero otherwise in order to reduce the number of elements in the spatial matrix to model.[10]

Group heterogeneity: In some cases, it is plausible to assume that spillovers are specific to pairs of groups of firms (see, e.g., Manresa (2013, p. 5.2.1)). Group heterogeneity implicitly assumes that firms within the same group (e.g. by industrial classification) are equally impacted by R&D of firms in other groups. Group heterogeneity is particularly helpful in dealing with unbalanced panel data with thousands of firms and few time periods. The

[10]See Lychagin, Pinkse, Slade, and Reenen (2016, Sect. IV) a discussion on this. The binary metric can be extended to capture proximity in other settings like industry, product market among others.

conditional expectation eq. (3.1), under group heterogeneity, is

$$E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) = \exp(\rho_0 + \rho_1 x_{it} + \rho_2 \sum_{L \not\ni i, K \ni i} w_{KLt}(\boldsymbol{\delta}) \sum_{j \in L} x_{jt} + \mathbf{z}_{it}\boldsymbol{\gamma}) \tag{4.1}$$

where $w_{KLt}(\boldsymbol{\delta}) \equiv \dfrac{\exp(\boldsymbol{d}'_{KLt}\boldsymbol{\delta})}{\sum_{j \neq i}^{N} \exp(\boldsymbol{d}'_{KLt}\boldsymbol{\delta})}$ and $\boldsymbol{d}_{KLt} \in \boldsymbol{D}_t$ is specific to groups K and L. The baseline case obtains when there are $N-1$ groups. Group heterogeneity leaves group membership across time unrestricted as long as group membership is observed. Formulation 4.1 has a major advantage of not allowing the dimension of $\boldsymbol{w}_t(\boldsymbol{\delta})$ to remain fixed as T grows *ad infinitum*.

5 Estimation

The model corresponding to eq. (3.1) cannot be estimated in two steps because of an intrinsic latency. Since $\boldsymbol{\delta}$ is an unknown parameter vector, the $N \times N$ spatial matrices $\boldsymbol{w}_t(\boldsymbol{\delta})$, $t = 1, \ldots, T$ are unknown ex-ante. The foregoing suggests a joint determination of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Estimation of eq. (3.4) is fairly straightforward using built-in optimisers in available software. Proceeding thus, however, raises a number of issues. First, not all models have smooth objective functions $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})$, e.g., quantile regression. Second, where the dimensionality of $\boldsymbol{\beta}$ is high (e.g., firm and year fixed effects with large N and T), direct minimisation of $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ becomes very slow and risks getting trapped in local minima. Third, direct optimisation fails to use efficient available routines for common models like the negative binomial, linear, logit, or Poisson regression. We propose the following estimation algorithm.

An Iterated Minimisation Scheme

In the following algorithm, we propose an iterative scheme that reduces direct optimisation to $\boldsymbol{\delta}$ while minimisation with respect to $\boldsymbol{\beta}$ becomes a standard regression problem that is easily handled using available routines.

Algorithm 1.

(a) Initialise counter $l = 0$ *and starting values* $\hat{\boldsymbol{\delta}}^{(l)}$

(b) Construct design matrix $\boldsymbol{m}^{(l)} = [\mathbf{1}_{NT}, \mathbf{x}, \boldsymbol{w}(\hat{\boldsymbol{\delta}}^{(l)})\mathbf{x}, \mathbf{z}]$ *and update counter* $l \leftarrow l + 1$

(c) Estimate $\hat{\boldsymbol{\beta}}^{(l)} = \underset{\boldsymbol{\beta} \in \boldsymbol{B}}{\arg\min}\, \mathcal{Q}_n(\boldsymbol{\beta}, \hat{\boldsymbol{\delta}}^{(l-1)})$, *i.e. regress* $\mathbf{y}$ *on* $\boldsymbol{m}^{(l)}$

(d) Solve $\hat{\boldsymbol{\delta}}^{(l)} = \underset{\boldsymbol{\delta} \in \boldsymbol{\Delta}}{\arg\min}\, \mathcal{Q}_n(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\delta})$

(e) If $\mathcal{Q}_n(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\delta}}^{(l-1)}) - \mathcal{Q}_n(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\delta}}^{(l)}) \leq \epsilon$, *stop else return to step (b)*

ϵ is a small user-specified number, e.g., $\epsilon = 10^{-6}$. Constructing the design matrix involves plugging $\hat{\boldsymbol{\delta}}^{(l)}$ into $\boldsymbol{w}(\cdot)$ with a pre-specified $\boldsymbol{D}$ and updating the column $\boldsymbol{w}(\hat{\boldsymbol{\delta}}^{(l)})\mathbf{x}$ in $\boldsymbol{m}^{(l)}$. The ease of algorithm 1 is at step (c) where regression of the outcome is run on the design matrix obtained in step (b) using available routines. The crux of Algorithm 1 reduces to the solving for $\hat{\boldsymbol{\delta}}^{(l)}$ at step (d) at each iteration.

6 Inference

This section concerns the asymptotic properties of our model. The asymptotic properties of concern are consistency and asymptotic normality. The treatment of consistency and asymptotic normality needs the concepts of conditionally strong-mixing processes and conditional stationarity. These are introduced in the following definitions.

Definition 1 (Strong conditional mixing)**.** *The sequence of random variables* $\{y_{it}\}_{t=1}^{T}$ *is conditionally strong-mixing ($\mathcal{F}$-strong-mixing) if there exists a non-negative $\mathcal{F}$-measurable random variable* $\alpha_i^{\mathcal{F}}(\iota)$ *converging to zero almost surely (a.s.) as* $\iota \to \infty$ *such that* $|\mathcal{P}(A_i \cap B_i|\mathcal{F}) - \mathcal{P}(A_i|\mathcal{F})\mathcal{P}(B_i|\mathcal{F})| \leq \alpha_i^{\mathcal{F}}(\iota)$ *for all* $A_i \in \sigma(y_{i1}, \ldots, y_{ik})$, $B_i \in \sigma(y_{ik+\iota}, \ldots, y_{iT})$, $1 \leq k \leq (T-1)$, $1 \leq \iota \leq (T-k)$, *and* $i \in \{1, \ldots, N\}$.

Definition 2 (Conditional Stationarity)**.** *The sequence of random variables* $\{y_{it}\}_{t=1}^{T}$ *for each* $i = 1, \ldots, N$ *is conditionally stationary if the joint distribution of* $\{y_{i\tau}\}_{\tau=t_1}^{t_k}$ *conditioned on*

$\mathcal{F}$ is the same as the joint distribution of $\{y_{i\tau}\}_{\tau=t_1+r}^{t_k+\iota}$ conditioned on $\mathcal{F}$ almost surely for all $1 \le t_k \le (T-1)$, $1 \le r \le (T-t_k)$.

The above definitions adapt definitions 4 and 5 in Rao (2009) to our context. The following assumptions are introduced.

Assumption 4. *(a) $E(y_{it}y_{j\tau}|\mathcal{F}) = E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t)E(y_{j\tau}|\mathbf{x}_\tau, \mathbf{z}_{j\tau}, \boldsymbol{D}_\tau)$ for any $i, j \in \{1, \ldots, N\}$, $i \neq j$, and $t, \tau \in \{1, \ldots, T\}$. (b) There exists a positive constant $\bar{\alpha}$, $\bar{\alpha}^8 < \infty$, such that for each $i \in \{1, \ldots, N\}$, the collection of conditionally strong-mixing random variables $\{y_{it}\}_{t=1}^T$ has an $\mathcal{F}$-strong-mixing coefficient that satisfies $\lim_{T\to\infty} \max_{i\in\{1,\ldots,N\}} \sum_{\iota=1}^{T} \alpha_i^{\mathcal{F}}(\iota) \le \bar{\alpha}$. (c) For each $i \in \{1, \ldots, N\}$, $\{y_{it}\}_{t=1}^T$ conditional on $\mathcal{F}$ is stationary.*

Note that assumption 4 does not assume serial independence for each y_{it}, $t = 1, \ldots, T$. Also, it does not impose unconditional independence between the outcomes y_{it} and $y_{j\tau}$ but requires conditional independence. The concepts of conditional independence and conditional stationarity allow us to establish a bound on the sum covariance of all random variable pairs in $\{y_{it}\}_{i=1,t=1}^{N,T}$ conditional on $\mathcal{F}$. The bound is established in the following proposition.

Proposition 1. *Under assumptions 3(b) and 4(a-c), $\lim_{N,T\to\infty} \frac{1}{NT}\Big|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{t=1}^{T}\sum_{\tau\neq t} \operatorname{cov}(y_{it}, y_{j\tau}|\mathcal{F})\Big| \le 8\bar{\alpha}\kappa_y$ a.s. where the conditional covariance $\operatorname{cov}(y_{it}, y_{j\tau}|\mathcal{F}) \equiv E(y_{it}y_{j\tau}|\mathcal{F}) - E(y_{it}|\mathcal{F})E(y_{j\tau}|\mathcal{F})$*

6.1 Consistency

To proceed with the proof of consistency, let us make the following assumption.

Assumption 5. *The parameter space $\boldsymbol{\Theta}$ of $\boldsymbol{\theta} = [\boldsymbol{\beta}', \boldsymbol{\delta}']'$ is a compact subset of $\mathbb{R}^{k_\theta}$.*

In the following lemma, we show that $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})$ converges uniformly in probability to its limit $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})$.

Lemma 5. *Under the assumptions of lemma 3 and assumption 4(a-c), $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})$ is continuous, and $\sup_{\substack{\boldsymbol{\beta}\in\boldsymbol{B}\\ \boldsymbol{\delta}\in\boldsymbol{\Delta}}} |\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta}) - \mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})| \xrightarrow{p} 0$.*

With the uniform convergence in probability result (lemma 5) in hand, the consistency of $\hat{\boldsymbol{\theta}} \equiv [\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}']'$ is stated in the following theorem.

Theorem 2 (Consistency)**.** *Under the assumptions of lemmas 1 to 5,* $[\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}']' \xrightarrow{p} [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ *where* $\boldsymbol{\theta}_o \equiv [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ *is the true parameter.*

6.2 Asymptotic Normality

Results on asymptotic normality require the score function and hessian matrix of $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})$. The score function is given by $s_n(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})$ where $s_{it}(\boldsymbol{\beta}, \boldsymbol{\delta}) = [\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}'}, \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}'}]'$ is $k_\theta \times 1$, $\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}} = \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) - y_{it})}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})} m_{it}(\boldsymbol{\delta})'$, and $\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) - y_{it})}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})} \frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}}$. The hessian matrix is given by $H_n(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})$ where

$$H_{it}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \begin{bmatrix} \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} & \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'} \\ \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\beta}'} & \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'} \end{bmatrix},$$

$$\frac{\partial q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \frac{\eta^{-2}(\eta^{-2} + y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1 + \eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2} m_{it}(\boldsymbol{\delta})' m_{it}(\boldsymbol{\delta}),$$
$$\frac{\partial q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'} = \frac{\eta^{-2}(\eta^{-2} + y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1 + \eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2} m_{it}(\boldsymbol{\delta})' \frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'} + \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) - y_{it})}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})} \frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'},$$
$$\frac{\partial q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'} = \frac{\eta^{-2}(\eta^{-2} + y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1 + \eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2} \frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}} \frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'} + \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) - y_{it})}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})} \frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'},$$

$\frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'} = [\mathbf{0}_{k_\beta \times 2}, \sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}, \mathbf{0}_{k_\beta \times (k_\beta - 2)}]'$, $\frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}} = \rho_2 \sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}$, $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = w_{ijt}(\boldsymbol{\delta})[\boldsymbol{d}_{ijt} - \sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}]$, $\frac{\partial (m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'} = \rho_2 \sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'}$, $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'} = [\boldsymbol{d}_{ijt} - \sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}]\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}'} - w_{ijt}(\boldsymbol{\delta})\sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}\boldsymbol{d}_{ilt}'$, and $\mathbf{0}_{a\times b}$ denotes an $a \times b$ zero matrix.

The proof of asymptotic normality requires conditions that are established in the following lemmata.

Lemma 6. *Under the assumptions of lemma 2, proposition 1, and assumption 3, (a)* $E[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] = \mathbf{0}$ *(b) All elements in the second moment of* $(NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$, $\mathrm{var}[(NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$, *are finite.*

Lemma 7. *Under the assumptions of lemma 2 and assumption 3, each element of the* $k_\theta \times k_\theta$ *hessian matrix* $H_n(\boldsymbol{\beta}, \boldsymbol{\delta})$ *is bounded in absolute value for all* $\boldsymbol{\beta} \in \boldsymbol{B}$ *and* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$.

The next result establishes the positive definiteness of $E[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$. The following definition and assumption are essential to establishing the result.

Definition 3. *For each* $i \in \{1, \ldots, N\}$ *and* $t \in \{1, \ldots, T\}$, *let* $\mathbf{a}^{it} = [a_1^{it}, \ldots, a_{i-1}^{it}, 0, a_{i+1}^{it}, \ldots, a_N^{it}]$ *be an arbitrary collection of random variables that satisfy* $0 < ||\mathbf{a}^{it}||_2 < \infty$ *a.s. Let* $\tilde{\boldsymbol{d}}_{i\cdot t}$ *be an* $N \times k_\delta$ *matrix whose* j*'th row is* $\boldsymbol{d}'_{ijt}$ *if* $j \neq i$ *and* $\mathbf{0}_{1\times k_\delta}$ *otherwise.*

Assumption 6. *(a) Any* $NT \times k_\delta$ *matrix* $\tilde{\boldsymbol{\mu}}$ *whose* $((t-1)N + i)$*'th row is* $\mathbf{a}^{it}\tilde{\boldsymbol{d}}_{i\cdot t}$ *such that* $\mathbf{a}^{it}$ *and* $\tilde{\boldsymbol{d}}_{i\cdot t}$ *are as in definition 3 for all* $i = 1, \ldots, T$ *and* $t = 1, \ldots, T$ *is full column rank. (b)* $\rho_{o,2} \neq 0$ *where* $\rho_{o,2}$ *is the third element in* $\boldsymbol{\beta}_o$.

Assumption 6 concerns the part of the hessian matrix corresponding to $\boldsymbol{\delta}$. It ensures non-singularity in the hessian matrix.

Lemma 8. *Under the assumptions of lemma 2, and assumption 6,* $E[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$ *is positive definite.*

In addition to the preceding lemmata, we impose the following standard assumption for the proof of asymptotic normality.

Assumption 7. $\boldsymbol{\theta}_o$ *is in the interior of* $\boldsymbol{\Theta}$

The following theorem provides the asymptotic normality result.

Theorem 3 (Asymptotic Normality)**.** *Under the assumptions of theorem 2, lemmata 6 to 8 and assumption 7,* $\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_\theta)$ *where* $\mathbf{V}_\theta$ *is the variance-covariance matrix.*

Parameters of interest - continued: Inference on parameters of interest viz. private, spillover, and social effects is important from a policy analysis perspective. Due to considerations of space, we consider averages of the parameters of interest: average private effect, $\varrho_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta}) = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})\boldsymbol{e}_2'\boldsymbol{\beta}$, the average spillover effect, $\varrho_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta}) = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j\neq i} w_{ijt}(\boldsymbol{\delta})\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})\boldsymbol{e}_3'\boldsymbol{\beta}$, and the average social effect $\varrho_{n,\mathcal{SE}}(\boldsymbol{\beta},\boldsymbol{\delta}) = \varrho_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta}) + \varrho_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta})$ where $\boldsymbol{e}_k$ denotes a $k_\beta \times 1$ matrix of zeros whose k'th element is 1. Asymptotic normality of the aforementioned parameters of interest require Jacobian matrices. The Jacobian matrices are given by $\varrho_{n,\mathcal{PE}}'(\boldsymbol{\beta},\boldsymbol{\delta}) = [\frac{\partial\varrho_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}'}, \frac{\partial\varrho_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}]'$ where $\frac{\partial\varrho_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})(\boldsymbol{e}_2 + \boldsymbol{e}_2'\boldsymbol{\beta}m_{it}(\boldsymbol{\delta})')$ and $\frac{\partial\varrho_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})\boldsymbol{e}_2'\boldsymbol{\beta}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}$, $\varrho_{n,\mathcal{S}p\mathcal{E}}'(\boldsymbol{\beta},\boldsymbol{\delta}) = [\frac{\partial\varrho_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}'}, \frac{\partial\varrho_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}]'$ where $\frac{\partial\varrho_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j\neq i} w_{ijt}(\boldsymbol{\delta})\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})(\boldsymbol{e}_3 + \boldsymbol{e}_3'\boldsymbol{\beta}m_{it}(\boldsymbol{\delta})')$ and $\frac{\partial\varrho_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j\neq i} w_{ijt}(\boldsymbol{\delta})\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})\boldsymbol{e}_3'\boldsymbol{\beta}(\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} + \frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}})$. The Jacobian matrix corresponding to the social effect $\varrho_{n,\mathcal{SE}}'(\boldsymbol{\beta},\boldsymbol{\delta}) = \varrho_{n,\mathcal{PE}}'(\boldsymbol{\beta},\boldsymbol{\delta}) + \varrho_{n,\mathcal{S}p\mathcal{E}}'(\boldsymbol{\beta},\boldsymbol{\delta})$.

The following corollary shows asymptotic normality of the parameters of interest.

Corollary 1. *Let $\varrho_n : \boldsymbol{\Theta} \to \mathbb{R}$ and $\varrho(\boldsymbol{\theta}) \equiv \lim_{N,T\to\infty}\varrho_n(\boldsymbol{\theta})$ denote a parameter of interest. Under assumption 7 and assumptions of theorem 3, $\sqrt{NT}(\varrho_n(\hat{\boldsymbol{\theta}}) - \varrho(\boldsymbol{\theta}_o)) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \varrho'(\boldsymbol{\theta}_o)'\mathbf{V}_\theta\varrho'(\boldsymbol{\theta}_o))$ where $\varrho'(\cdot)$ denotes a $k_\theta \times 1$ the Jacobian matrix of a parameter of interest.*

7 Empirical Application

Quantifying knowledge spillovers of R&D is an important research pursuit because of its policy implications. In this section, we estimate private and spillover effects of R&D on firm innovation. This problem is studied in Bloom, Schankerman, and Van Reenen (2013), Cincera (1997), and Grillitsch and Nilsson (2015).

7.1 Data

Data for our empirical analyses are an updated version of data used in Bloom, Schankerman, and Van Reenen (2013).[11] A balanced panel is constructed out of the original data set

[11] The updated version is used in Lucking, Bloom, and Van Reenen (2018). The data and codes are accessible at `https://nbloom.people.stanford.edu/research`.

hence we do not replicate the exact results in Bloom, Schankerman, and Van Reenen (2013). The balanced panel comprises 217 firms in 21 two-digit SIC industries from 1985 to 2011. Table 1 presents the number of firms by 2-digit industry classification (SIC2). Prominent among the industries are the Electronic & Other Electric Equipment (38 firms), Chemicals and Allied Products (34 firms), Industrial Machinery & Equipment (33 firms), and Instruments & Related Products (31 firms). Majority of the firms (95%) are in the manufacturing sector.

Table 1: Firms by 2-digit SIC

SIC2	**Industry**	**no. of firms**
13	Oil and Gas Extraction	2
14	Nonmetalic minerals, except fuels	1
20	Food and Kindred Products	6
24	Lumber and Wood Products	1
25	Furniture and Fixtures	7
26	Paper and Allied Products	5
27	Printing and Publishing	1
28	Chemicals and Allied Products	34
29	Petroleum and Coal Products	1
30	Rubber & Misc. Plastics Products	4
32	Stone, Clay, and Glass Products	3
33	Primary Metal Industries	4
34	Fabricated Metal Products	13
35	Industrial Machinery & Equipment	33
36	Electronic & Other Electric Equipment	38
37	Transport Equipment	23
38	Instruments & Related Products	31
39	Misc. Manufacturing Industries	3
50	Wholesale Trade - Durable Goods	1
73	Business Services	5
99	Unclassified	1

Table 2 presents summary statistics of the main variables. The outcome variable is citation-weighted patent counts Pat_Cite. As Aghion, Van Reenen, and Zingales (2013) argued, citation-weighted patent counts not only capture the quantity of R&D output (patent counts) but also the acknowledged relevance (citations). Pat_Cite has a mass point at zero

Table 2: Summary Statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
Pat_Cite	0	0	6	453.641	158	5714	1218.84
Patent flow	0	0	3	80.113	30	4344	267.668
R&D stock	0	37.513	181.609	2253.656	1302.627	47296.664	5663.882
$ln(SpTECH)$	3.649	5.538	5.919	5.95	6.305	7.933	0.65
$ln(SpSIC)$	4.813	5.193	5.495	5.449	5.709	5.977	0.308
$ln(SpGEOG)$	1.625	4.748	5.232	5.364	5.757	9.256	1.005
Sales	0.827	356.386	1514.999	11131.897	6748.933	1176922.25	51194.824

Notes: All values above are based on firm-year observations. The outcome variable is citation-weighted patent counts (Pat_Cite). Other variables include patent counts (Patent flow), stock of R&D (R&D stock), spillover measure using technology proximity $ln(SpTECH)$, spillover measure using product market proximity $ln(SpSIC)$, spillover measure using geographical proximity $ln(SpGEOG)$, real sales revenue (Sales). $ln(SpTECH)$, $ln(SpSIC)$, and $ln(SpGEOG)$ are constructed à la Bloom, Schankerman, and Van Reenen (2013) using the Jaffe (1986) measures related to technology, product market, and geographic distances with the log of lagged R&D stock for the 217 firms in the balanced panel. Prices are measured in millions of 1996 dollars.

as about 43% of the observations have zero citation-weighted patent counts. Patent counts (Patent flow) is zero for about 38% of the firm-year observations. The main covariate of interest is R&D (R&D stock). Bloom, Schankerman, and Van Reenen (2013)'s exogenously constructed measures of spillovers, $ln(SpTECH)$ (in technology space) and $ln(SpSIC)$ (in product market space), and $ln(SpGEOG)$ (using geographical location) are included for comparison. To control for firm size, we include log real sales revenue[12] $ln(Sales)$. Pre-sample mean scaling approach $Pre-SFE$, which conditions on pre-sample citation-weighted patents, is included in our specifications to estimate fixed effects à la Blundell, Griffith, and Van Reenen (1999). We include 3-digit SIC industry and year dummies in all specifications to control for unobserved heterogeneity.

7.2 Empirical model

We consider the negative binomial model because the outcome variable (citation-weighted patent counts) is discrete. While there exist alternatives to the negative binomial model for

[12]Nominal sales are deflated by industry price indices to obtain real sales.

handling outcome variables with mass points at zero (see discussion in Berger, Stocker, and Zeileis (2017) and Huang and Tsyawo (2018)), the crux of our empirical application lies in eliciting spatial interactions and estimating spillover effects in a possibly non-linear model. Our model is applicable to a wide class of models which admit linear predictor functions, and since this class encompasses a number of methods used in the literature viz. zero-inflated negative binomial, zero-inflated Poisson, hurdle models, Tobit, Heckman selection, logit, distribution regression, and quantile regression, our approach remains applicable.

The conditional mean is given by

$$E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t) = \exp(\rho_0 + \rho_1 x_{it} + \rho_2 \sum_{j \neq i} x_{jt} w_{ijt}(\boldsymbol{\delta}) + \mathbf{z}_{it}\boldsymbol{\gamma}) \tag{7.1}$$

where the exponential link function is used for the negative binomial model, ρ_1 and ρ_2 respectively denote the private and spillover elasticities of research and development ($R\&D$) on firm innovation. x_{it} denotes firm i's R&D (lagged log R&D stock) and $\mathbf{z}_{it}$ contains firm characteristics viz. lagged log real sales, a dummy variable for observations with zero lagged R&D stock, a dummy for observations with lagged patent stock equal to zero, a "pre-sample mean-scaling approach" to estimate fixed effects à la Blundell, Griffith, and Van Reenen (1999), 3-digit industry dummies and year dummies. In some specifications, spillover measures $ln(SpTECH)$, $ln(SpSIC)$, and $ln(SpGEOG)$ are included. Standard errors are clustered at the firm level.

Spillover covariates in $\boldsymbol{d}_{ijt} \in \boldsymbol{D}_t$ considered are the Jaffe measures in technology ($dTECH$), product market ($dSIC$), and geographical proximity ($dGEOG$), 6 bivariate polynomial expansion terms of $R\&D_{i,t-2}, (dR\&D_{j,t-2})$, $j \neq i$ up to the third order, a dummy equal to one if firms i, j are in the same 3-digit industry classification ($dI(SIC3)$), a dummy equal to one if one (but not both firms) has zero lagged R&D stock $dI(R\&D)$, and a dummy equal to one if one (but not both firms) has a lagged patent stock equal to zero $dI(Cite)$. The polynomial series terms on $R\&D_{i,t-2}$ and $R\&D_{j,t-2}$ is given by a vector comprising

$\{R\&D^{l_i}_{i,t-2} R\&D^{l_j}_{i,t-2} : l_i = 0, \ldots, k-1, l_j = 1, \ldots, k-l_i\}$ and $k = 3$.

7.3 Results

The main results are summarised in tables 3 and 4. Since the main covariate $\mathbf{x}$ of interest is in logarithms, the coefficients ρ_1 and ρ_2 are interpretable as elasticities. The coefficient ρ_1 on $ln(R\&D)_{t-1}$ denotes the elasticity of own R&D to innovation whereas spillover elasticity of R&D using our model is captured by the coefficient ρ_2. Bloom, Schankerman, and Van Reenen (2013)'s exogenously constructed spillover measures are included in specifications (1) and (7) in table 3 and in specification (5) in table 4. Columns (2)-(7) in table 3 and columns (1)-(5) in table 4 are specifications using our approach with different combinations of spatial covariates $\boldsymbol{d}_{ijt}$ in the weight function $w_{ijt}(\boldsymbol{\delta})$. The main specifications of interest are columns (6) and (7) in table 3; these combine all spatial covariates in the weight function and allows us to assess all spatial covariates simultaneously.

In each table, panel A reports coefficients with standard errors in parentheses. The standard errors are clustered by firm to allow serial correlation. Panel B reports Wald-statistics (chi-square statistic) corresponding to the spatial covariate (subscript to χ^2), the degree of freedom df, and the p-value $P(> \chi^2)$ of the chi-square statistic (in parentheses).

Results comparable to tables 3 and 4 can be found in Bloom, Schankerman, and Van Reenen (2013, table IV) and Lucking, Bloom, and Van Reenen (2018, table 3). Specification (1) includes Bloom, Schankerman, and Van Reenen (2013)'s exogenously constructed spillover measures, namely, $ln(SpTECH)_{t-1}$, $ln(SpSIC)_{t-1}$, and $ln(SpGEOG)_{t-1}$ using the balanced panel. Out of the three spillover measures, $ln(SpGEOG)_{t-1}$ is positive and significant at the 5% level whereas $ln(SpTECH)_{t-1}$ and $ln(SpSIC)_{t-1}$ are not significant at any of the conventional levels. One makes a similar observation in specification (7) of table 3 and specification (5) of table 4 where the coefficient on $ln(SpGEOG)$ is significant at the 1% level (in both specifications). Observe that the coefficient on $ln(SpTECH)$ is negative (but not significant) in specification (7) of table 3 while the coefficient on $ln(SpSIC)$ is

Table 3: Coefficients - Citation-Weighted Patent Counts I

	Neg. Bin (1)	Neg. Bin (2)	Neg. Bin (3)	Neg. Bin (4)	Neg. Bin (5)	Neg. Bin (6)	Neg. Bin (7)
Panel A: ($\boldsymbol{\beta}$)							
ρ_1	0.155	0.187	0.176	0.186	0.181	0.170	0.157
	(0.044)	(0.042)	(0.043)	(0.042)	(0.041)	(0.041)	(0.045)
ρ_2		0.048	-1.331	1.566	0.213	0.319	0.325
		(0.402)	(1.173)	(0.534)	(0.070)	(0.067)	(0.067)
$ln(SpTECH)_{t-1}$	0.059						-0.016
	(0.126)						(0.124)
$ln(SpSIC)_{t-1}$	-2.818						-0.939
	(2.084)						(2.225)
$ln(SpGEOG)_{t-1}$	0.068						0.085
	(0.029)						(0.031)
$ln(Pat)_{t-1}$	0.551	0.560	0.555	0.552	0.584	0.563	0.559
	(0.034)	(0.035)	(0.033)	(0.035)	(0.034)	(0.033)	(0.033)
$Pre-SFE$	0.172	0.147	0.153	0.164	0.164	0.180	0.195
	(0.038)	(0.038)	(0.039)	(0.040)	(0.039)	(0.039)	(0.040)
Panel B: ($\boldsymbol{\delta}$)							
χ^2_{dTECH}		0.210				0.94	1.000
df, $P(>\chi^2)$		1,(0.650)				1,(0.330)	1,(0.310)
χ^2_{dSIC}			122.200			10.600	13.400
df, $P(>\chi^2)$			1,(0.000)			1,(0.001)	1,(0.000)
χ^2_{dGEOG}				34.100		14.000	12.000
df, $P(>\chi^2)$				1,(0.000)		1,(0.000)	1,(0.001)
$\chi^2_{dR\&D_{t-2}}$					190.700	50.000	49.100
df, $P(>\chi^2)$					6,(0.000)	6,(0.000)	6,(0.000)
$\chi^2_{dI(SIC3)}$						1.000	0.910
df, $P(>\chi^2)$						1,(0.310)	1, (0.340)
$\chi^2_{dI(R\&D)}$						0.76	0.620
df, $P(>\chi^2)$						1,(0.380)	1,(0.430)
$\chi^2_{dI(Cite)}$						5.000	5.200
df, $P(>\chi^2)$						1,(0.025)	1,(0.022)

Notes: All specifications above (Neg. Bin (1)-(5)) contain firm and year fixed effects. The outcome variable is citation-weighted patent counts (Pat_Cite_t). Number of firm-year observations: 5859. The spillover-generating variable $\mathbf{x}$ is lagged log R&D $ln(R\&D)_{t-1}$. Spillover covariates (with respective Wald-statistics) considered: Jaffe measures in technology ($dTECH$), product market ($dSIC$), and geographical proximity ($dGEOG$), 6 bivariate polynomial expansion terms of $R\&D_{i,t-2}, (dR\&D_{j,t-2}),\ j \neq i$ up to the third order, a dummy equal to one if firms i, j are in the same 3-digit industry classification ($dI(SIC3)$), a dummy equal to one if one (but not both firms) has zero lagged R&D stock $dI(R\&D)$, and a dummy equal to one if one (but not both firms) has a lagged patent stock equal to zero $dI(Cite)$. $P(>\chi^2)$ denotes the p-value and df equals the length of the corresponding parameter vector.

consistently negative.

Considering results in tables 3 and 4, one observes that the elasticity of own R&D on innovation is positive and significant at the 1% level across all model specifications. Relative to specification (1), the elasticity of own R&D (ρ_1) in other specifications (where our approach is used) is slightly higher in magnitude save specification (5) in table 4. The corresponding standard errors do not change much across specifications in both tables.

The coefficient ρ_2 in specifications (2) through (5) in table 3 and specifications (2)-(5) in table 4 gives the spillover elasticity of $R\&D$ on firm innovation using our model. Notice that the sign and magnitude vary by spatial covariate(s) included in the weight function. ρ_2 is negative in specifications (3) of table 3 and (2)-(5) of table 4 where $dSIC$, $dI(SIC3)$, $dI(R\&D)$, and $dI(Cite)$ are spatial covariates included in the weight function. $dTECH$, $dGEOG$, $dR\&D_{t-2}$ in specifications (2), (3)-(5) spatial covariates associated with positive ρ_2. Save specifications (2)-(3) in table 3 and (2)-(3) in table 4, ρ_2 is statistically significant at the 1% level across specifications. One also notes variation in the standard errors of ρ_2 by the spatial covariate included in the weight function. The significance of ρ_2 is also indicative of the presence of spillovers and cross-sectional dependence in firm patenting behaviour. Coefficients on the lagged outcome variable $ln(Pat)_{t-1}$ are positive and significant at the 1% level across all specifications which confirms strong persistence in firm patenting behaviour as in Bloom, Schankerman, and Van Reenen (2013) and Lucking, Bloom, and Van Reenen (2018).

The significance of coefficients $\boldsymbol{\delta}$ on the spatial covariates $\boldsymbol{d}_{ijt}$ is indicative of support against the null hypothesis that each element of $\boldsymbol{w}(\boldsymbol{\delta})$ is equal to $1/(N-1)$ where N is the number of firms.[13] Rejecting this hypothesis does, however, not indicate the relevance of the spatial covariate as long as the outcome is concerned. It ought to be coupled with ρ_2 for the relevance of the spatial covariate to be confirmed by the data. In this vein, one notes that while $\boldsymbol{\delta}$ on $dSIC$ in table 3(3), $dI(SIC3)$ in table 4(2), and $dI(R\&D)$ table 4(3)

[13] $w_{ijt}(\boldsymbol{\delta}) \equiv \dfrac{\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\sum_{j\neq i}^{N}\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})} = 1/(N-1)$ if $\boldsymbol{\delta} = \boldsymbol{0}$, $\forall\ i = 1, \ldots, N,\ t = 1, \ldots, T$

Table 4: Coefficients - Citation-Weighted Patent Counts II

	Neg. Bin (1)	Neg. Bin (2)	Neg. Bin (3)	Neg. Bin (4)	Neg. Bin (5)
Panel A: ($\boldsymbol{\beta}$)					
ρ_1	0.141	0.167	0.187	0.175	0.151
	(0.045)	(0.046)	(0.042)	(0.042)	(0.045)
ρ_2	-0.235	-1.518	-0.056	-0.227	-0.228
	(0.045)	(1.169)	(0.191)	(0.047)	(0.046)
$ln(SpTECH)_{t-1}$	0.044				0.053
	(0.132)				(0.129)
$ln(SpSIC)_{t-1}$	-0.567				-0.993
	(2.328)				(2.290)
$ln(SpGEOG)_{t-1}$	0.077				0.077
	(0.029)				(0.028)
$ln(Pat)_{t-1}$	0.556	0.561	0.560	0.557	0.550
	(0.034)	(0.035)	(0.035)	(0.033)	(0.033)
$Pre-SFE$	0.151	0.143	0.147	0.132	0.147
	(0.036)	(0.039)	(0.039)	(0.036)	(0.036)
Panel B: ($\boldsymbol{\delta}$)					
$\chi^2_{dI(SIC3)}$	12.400	92.500			
df, $P(>\chi^2)$	1,(0.000)	1,(0.000)			
$\chi^2_{dI(R\&D)}$	81.100		17.400		
df, $P(>\chi^2)$	1,(0.000)		1,(0.000)		
$\chi^2_{dI(Cite)}$	9.200			19.000	21.600
df, $P(>\chi^2)$	1,(0.002)			1,(0.000)	1,(0.000)

Notes: All specifications above (Neg. Bin (1)-(5)) contain firm and year fixed effects. The outcome variable is citation-weighted patent counts (Pat_Cite_t). Number of firm-year observations: 5859. The spillover-generating variable $\mathbf{x}$ is lagged log R&D $ln(R\&D)_{t-1}$. Spillover covariates (with respective Wald-statistics) considered: a dummy equal to one if firms i, j are in the same 3-digit industry classification ($dI(SIC3)$), a dummy equal to one if one (but not both firms) has zero lagged R&D stock $dI(R\&D)$, and a dummy equal to one if one (but not both firms) has a lagged patent stock equal to zero $dI(Cite)$. $P(>\chi^2)$ denotes the p-value and df equals the length of the corresponding parameter vector.

are statistically significant at the 1% level, the corresponding ρ_2 are not significant at any conventional level.

Specifications (6) and (7) in table 3 allows a comparison of all spatial covariates simultaneously. One notes that elements of $\boldsymbol{\delta}$ corresponding to $dTECH$, $dSIC3$,, and $dI(R\&D)$ are not significant. Those on $dSIC$, $dGEOG$, $dR\&D_{t-2}$ are significant at the 1% level while the coefficient in $\boldsymbol{\delta}$ on $dI(Cite)$ is significant at the 5% level. ρ_2 in specifications (6) and (7) are positive and significant at the 1% level, confirming that R&D spillover effects on firm innovation are positive and the relevance of $dSIC$, $dGEOG$, $dR\&D_{t-2}$ and $dI(Cite)$ as spatial covariates. In the relevant set of spatial covariates, $dR\&D_{t-2}$ and $dI(Cite)$ are time-varying and $dR\&D_{t-2}$ is asymmetric.

In sum, our estimation (specification (6) in table 3) confirms positive private spillover effect of R&D on innovation. A 1% increase in own R&D is associated with a 0.170% increase in innovation. The private effect is dominated by the spillover effect of innovation. If every other firm increases R&D by 1%, innovation is expected to increase by 0.319%. In addition to the significance of the spillover parameter ρ_2, the relevant set of spatial covariates are associated with proximity in the product market space ($dSIC$), geographic proximity ($dGEOG$), past R&D ($dR\&D_{t-2}$), and past patent citations $dI(Cite)$. ($dR\&D_{t-2}$) and $dI(Cite)$ are not tied to any measure of proximity (e.g. geographic contiguity) and are time-varying.

8 Conclusion

The problem of quantifying R&D spillover effects on firm innovation among other outcomes crucially depends on the spatial matrix that is used to construct the R&D spillover measure. In this paper, we address the problem of unknown spatial matrices in a study of R&D spillover effects on firm innovation using the negative binomial model. The study extends to other studies that quantify spillover effects. As a key contribution of the paper, the proposed approach parsimoniously models and estimates the spatial matrix from panel data. The approach allows time variation and asymmetry in the spatial matrix and extends to a general class of linear and non-linear models that admit linear predictor functions. Our approach tackles the problem of metric ambiguity by accommodating several spillover covariates simultaneously and assesses their relevance via hypothesis tests. We provide identification results of the model and establish consistency and asymptotic normality of the estimator.

We apply our method to study R&D spillovers on innovation in a knowledge production framework. The importance of our approach is confirmed by the sensitivity of spillover effect estimates to the choice of spatial covariate and spatial matrix by extension. Our results confirm the presence of positive and statistically significant spillover effects of R&D on innovation. Though the private effects of R&D on innovation are positive and statistically significant, they are dominated by the spillover effects of R&D. In a specification that assesses the relevance of several spatial covariates, we find that product market proximity, geographic proximity, past R&D, and past patent citations are relevant sources of connectivity between firms with respect to innovation. The result on relevant spillover covariates confirm time-variation and asymmetry in the interaction structure between firms and that the strength of connectivity between firms is not only tied to commonly assumed notions of closeness viz. technological, product market, industry, and geographic proximity.

References

[1] Acemoglu, Daron, Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti. “Technology, Information, and the Decentralization of the Firm”. *The Quarterly Journal of Economics* 122.4 (2007), pp. 1759–1799.

[2] Aghion, Philippe, John Van Reenen, and Luigi Zingales. “Innovation and institutional ownership”. *American Economic Review* 103.1 (2013), pp. 277–304.

[3] Anselin, Luc, Julie Le Gallo, and Hubert Jayet. “Spatial panel econometrics”. *The econometrics of panel data*. Springer, 2008, pp. 625–660.

[4] Berger, Susanne, Herbert Stocker, and Achim Zeileis. “Innovation and institutional ownership revisited: an empirical investigation with count data models”. *Empirical Economics* 52.4 (2017), pp. 1675–1688.

[5] Bernstein, Jeffrey I and M Ishaq Nadiri. “Research and development and intra-industry spillovers: an empirical application of dynamic duality”. *The Review of Economic Studies* 56.2 (1989), pp. 249–267.

[6] Bhattacharjee, Arnab and Chris Jensen-Butler. “Estimation of the spatial weights matrix under structural constraints”. *Regional Science and Urban Economics* 43.4 (2013), pp. 617–634.

[7] Bloom, Nicholas, Mark Schankerman, and John Van Reenen. “Identifying technology spillovers and product market rivalry”. *Econometrica* 81.4 (2013), pp. 1347–1393.

[8] Blundell, Richard, Rachel Griffith, and John Van Reenen. “Market share, market value and innovation in a panel of British manufacturing firms”. *The Review of Economic Studies* 66.3 (1999), pp. 529–554.

[9] Boyd, Stephen and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[10] Case, Anne C. “Spatial patterns in household demand”. *Econometrica* 59.4 (1991), pp. 953–965.

[11] Cincera, Michele. “Patents, R&D, and technological spillovers at the firm level: some evidence from econometric count models for panel data”. *Journal of Applied Econometrics* 12.3 (1997), pp. 265–280.

[12] Coe, David T, Elhanan Helpman, and Alexander W Hoffmaister. “International R&D spillovers and institutions”. *European Economic Review* 53.7 (2009), pp. 723–741.

[13] Cohen, Wesley M and Daniel A Levinthal. “Innovation and learning: the two faces of R & D”. *The Economic Journal* 99.397 (1989), pp. 569–596.

[14] Griffith, Rachel, Stephen Redding, and John Van Reenen. “Mapping the two faces of R&D: Productivity growth in a panel of OECD industries”. *Review of Economics and Statistics* 86.4 (2004), pp. 883–895.

[15] Grillitsch, Markus and Magnus Nilsson. “Innovation in peripheral regions: Do collaborations compensate for a lack of local knowledge spillovers?” *The Annals of Regional Science* 54.1 (2015), pp. 299–321.

[16] Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg. “Market value and patent citations”. *RAND Journal of Economics* (2005), pp. 16–38.

[17] Hoshino, Tadao. “Semiparametric Estimation of Censored Spatial Autoregressive Models”. *Econometric Theory* (2017), pp. 1–38.

[18] Huang, Weige and Emmanuel Tsyawo. “Bayesian Distribution Regression”. *Available at SSRN 3048658* (2018).

[19] Jaffe, Adam B. *Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value*. 1986.

[20] Kapetanios, George, James Mitchell, and Yongcheol Shin. "A nonlinear panel data model of cross-sectional dependence". *Journal of Econometrics* 179.2 (2014), pp. 134–157.

[21] Kelejian, Harry H and Gianfranco Piras. "Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes". *Regional Science and Urban Economics* 46 (2014), pp. 140–149.

[22] Lam, Clifford and Pedro CL Souza. "Estimation and selection of spatial weight matrix in a spatial lag model". *Journal of Business & Economic Statistics* (2019), pp. 1–41.

[23] Lucking, Brian, Nicholas Bloom, and John Van Reenen. *Have R&D Spillovers Changed?* Tech. rep. National Bureau of Economic Research, 2018.

[24] Lychagin, Sergey, Joris Pinkse, Margaret E Slade, and John Van Reenen. "Spillovers in space: Does geography matter?" *The Journal of Industrial Economics* 64.2 (2016), pp. 295–335.

[25] Manresa, Elena. "Estimating the structure of social interactions using panel data". *Unpublished Manuscript. CEMFI, Madrid* (2013).

[26] Newey, Whitney K and Daniel McFadden. "Large sample estimation and hypothesis testing". *Handbook of Econometrics* 4 (1994), pp. 2111–2245.

[27] Pinkse, Joris, Margaret E Slade, and Craig Brett. "Spatial price competition: a semiparametric approach". *Econometrica* 70.3 (2002), pp. 1111–1153.

[28] Rao, BLS Prakasa. "Conditional independence, conditional mixing and conditional association". *Annals of the Institute of Statistical Mathematics* 61.2 (2009), pp. 441–460.

[29] Soale, Abdul-Nasah and Emmanuel Tsyawo. "Clustered Covariate Regression". *Available at SSRN: https://ssrn.com/abstract=3394012* (2019).

[30] Syverson, Chad. “What determines productivity?” *Journal of Economic Literature* 49.2 (2011), pp. 326–65.

[31] Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[32] Xu, Xingbai and Lung-fei Lee. “A spatial autoregressive model with a nonlinear transformation of the dependent variable”. *Journal of Econometrics* 186.1 (2015), pp. 1–18.

A Useful lemmata

In the following lemma, we provide bounds on derivative terms in the asymptotic variance-covariance matrix. Let $\kappa_\beta \equiv \sup_{\boldsymbol{\beta}\in\boldsymbol{B}} ||\boldsymbol{\beta}||_\infty$ and note that $\kappa_\beta < \infty$ by the compactness of $\boldsymbol{\Theta}$ (assumption 5).

Lemma 9. *Under the set of assumptions 3, the following hold (a)* $\sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \left|\left|\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}\right|\right|_2 \leq \frac{2}{N}\kappa_d\kappa_w$ *(b)* $\sup_{\substack{\boldsymbol{\beta}\in\boldsymbol{B}\\ \boldsymbol{\delta}\in\boldsymbol{\Delta}}} \left|\left|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\right|\right|_2 \leq 2\kappa_d\kappa_w\kappa_z$*, (c)* $\sup_{\substack{\boldsymbol{\beta}\in\boldsymbol{B}\\ \boldsymbol{\delta}\in\boldsymbol{\Delta}}} \left|\left|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\right|\right|_2 \leq 2\kappa_\beta\kappa_d\kappa_w\kappa_z$*, and (d)* $\sup_{\substack{\boldsymbol{\beta}\in\boldsymbol{B}\\ \boldsymbol{\delta}\in\boldsymbol{\Delta}}} \left|\left|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\right|\right|_2 \leq 5\kappa_\beta\kappa_d^2\kappa_w\kappa_z$

Proof of lemma 9. (a) Let us begin by bounding the term $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}$,

$$
\begin{aligned}
||\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}||_2 &= ||w_{ijt}(\boldsymbol{\delta})[\boldsymbol{d}_{ijt} - \sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}]||_2 \leq \frac{\kappa_w}{N}||\boldsymbol{d}_{ijt} - \sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}||_2 \\
&\leq \frac{\kappa_w}{N}||\boldsymbol{d}_{ijt}||_2 + \frac{\kappa_w}{N}||\sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}||_2 \leq \frac{\kappa_w}{N}||\boldsymbol{d}_{ijt}||_2 + \frac{\kappa_w}{N}\sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})||\boldsymbol{d}_{ilt}||_2 \leq \frac{2}{N}\kappa_d\kappa_w
\end{aligned}
$$

The first inequality follows from lemma 1. The conclusion follows from assumption 3(c).

(b) Using the bound on $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}$ from (a) above, and noting that only the non-zero elements matter in the norm, $||\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}||_2 = ||\sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}||_2 \leq \sum_{j\neq i}|x_{jt}|\cdot||\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}||_2 \leq 2\kappa_d\kappa_w\kappa_z$. The conclusion follows from assumption 3(c).

(c) $||\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}||_2 = ||\rho_2\sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}||_2 \leq |\rho_2|\cdot||\sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}||_2 \leq 2\kappa_\beta\kappa_d\kappa_w\kappa_z$. The conclusion follows from assumption 3(b) and using the result from (b) above.

(d) Bounding $\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}$ first requires a bound on $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}$.

$$
\begin{aligned}
||\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}||_2 &= ||[\boldsymbol{d}_{ijt} - \sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}]\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'} - w_{ijt}(\boldsymbol{\delta})\sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}\boldsymbol{d}_{ilt}'||_2 \\
&\leq ||[\boldsymbol{d}_{ijt} - \sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}]||_2||\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}||_2 + w_{ijt}(\boldsymbol{\delta})\sum_{l\neq i} w_{ilt}(\boldsymbol{\delta})||\boldsymbol{d}_{ilt}\boldsymbol{d}_{ilt}'||_2 \leq \frac{5}{N}\kappa_d^2\kappa_w
\end{aligned}
$$

With the bound on $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}$ in hand, $||\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}||_2 = ||\rho_2\sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}||_2 \leq |\rho_2|\sum_{j\neq i}|x_{jt}|\cdot$

$||\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'}||_2 \leq 5\kappa_\beta \kappa_d^2 \kappa_w \kappa_z$ □

B Proofs of lemmata and theorems

Proof of lemma 1.

First, we need to show that $Nw_{ijt}(\boldsymbol{\delta}) = \dfrac{N\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\sum_{j\neq i}^{N}\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}$ is bounded a.s. for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$. Note that

$$
\begin{aligned}
Nw_{ijt}(\boldsymbol{\delta}) &= \frac{\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{N^{-1}\sum_{j\neq i}^{N}\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})} \leq \frac{\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\exp(N^{-1}\sum_{j\neq i}^{N}\boldsymbol{d}'_{ijt}\boldsymbol{\delta})} \\
&= \exp((\boldsymbol{d}_{ijt} - \bar{\boldsymbol{d}}_{i.t})'\boldsymbol{\delta}) \leq \sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \exp((\boldsymbol{d}_{ijt} - \bar{\boldsymbol{d}}_{i.t})'\boldsymbol{\delta}) \leq \kappa_w
\end{aligned}
$$

The first inequality follows by Jensen's inequality since the $\exp(\cdot)$ is convex, and $N^{-1}\sum_{j\neq i}^{N}\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta}) \geq \exp(N^{-1}\sum_{j\neq i}^{N}\boldsymbol{d}'_{ijt}\boldsymbol{\delta})$. The last two inequalities follow from the sup operator and assumption 3(a).

Second, we know from the first part that $Nw_{ijt}(\boldsymbol{\delta}) \leq \kappa_w$ a.s. This implies that for $j \in \{1,\ldots,N\}$ and $t \in \{1,\ldots,T\}$,

$$
\sum_{i=1}^{N} w_{ijt}(\boldsymbol{\delta}) \leq \max_{\substack{j\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T\}}} \sum_{i=1}^{N} w_{ijt}(\boldsymbol{\delta}) \equiv ||\boldsymbol{w}(\boldsymbol{\delta})||_1 \leq N \sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \max_{\substack{j\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T\}\\ i\neq j}} w_{ijt}(\boldsymbol{\delta}) \leq N\frac{\kappa_w}{N} = \kappa_w \text{ a.s.}
$$

Note that by definition, $||\boldsymbol{w}(\boldsymbol{\delta})||_1 \equiv \max_{\substack{j\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T\}}} \sum_{i=1}^{N} w_{ijt}(\boldsymbol{\delta})$ because $\boldsymbol{w}(\boldsymbol{\delta})$ is block-diagonal.

□

Proof of lemma 2. Without loss of generality, and for ease of exposition, use the column-permuted $\boldsymbol{m}$ as $\boldsymbol{m} = [\boldsymbol{z},\ \boldsymbol{w}\mathbf{x}]$ and suppress dependence of $\boldsymbol{m}$ on $\boldsymbol{\delta}$. The Gram matrix of the design matrix $\boldsymbol{m} = [\boldsymbol{z},\ \boldsymbol{w}\mathbf{x}]$ is given by

$$
\boldsymbol{m}'\boldsymbol{m} = \begin{bmatrix} \boldsymbol{z}'\boldsymbol{z} & \boldsymbol{z}'\boldsymbol{w}\mathbf{x} \\ \mathbf{x}'\boldsymbol{w}'\boldsymbol{z} & \mathbf{x}'\boldsymbol{w}'\boldsymbol{w}\mathbf{x} \end{bmatrix}
$$

Note that under assumption 2(a), $\boldsymbol{z}'\boldsymbol{z}$ is invertible. By assumption 2(b), the block diagonal matrix $\boldsymbol{w}$ is full rank. Using the result in Boyd and Vandenberghe (2004, sect A.5.5) on the positive definiteness of symmetry matrices, we only need to show that the Shur complement of $\boldsymbol{z}'\boldsymbol{z}$ in $\boldsymbol{m}'\boldsymbol{m}$, $S = \mathbf{x}'\boldsymbol{w}'\boldsymbol{w}\mathbf{x} - \mathbf{x}'\boldsymbol{w}'\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{w}\mathbf{x} > 0$.

Factor S as $S = \mathbf{x}'\boldsymbol{w}'[\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}']\boldsymbol{w}\mathbf{x}$. Note that $\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'$ is a projection matrix, i.e., it is symmetric and idempotent; $[\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}']$ is also symmetric and idempotent. Since $\mathbf{x} \neq \mathbf{0}$ a.s. (assumption 2(a)), it implies S can be expressed as $S = \mathbf{x}'\boldsymbol{w}'[\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'][\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}']\boldsymbol{w}\mathbf{x} = \sum_{i=1}^{N}\sum_{t=1}^{T}\zeta_{it}^2 > 0$ where $\zeta_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T$ are residuals obtained by regressing $\boldsymbol{w}\mathbf{x}$ on $\boldsymbol{z}$, i.e., S is the sum of squared residuals. From the foregoing, $\frac{1}{NT}\boldsymbol{m}'\boldsymbol{m}$ is positive definite.

Then from assumption 3(b), $\sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} ||m_{it}(\boldsymbol{\delta})||_2 \leq \kappa_m$ a.s. It follows from the foregoing that $\frac{1}{NT}||\boldsymbol{m}'\boldsymbol{m}||_2 \leq \frac{1}{NT}||\boldsymbol{m}||_2^2 \leq \kappa_m^2 < \infty$ a.s. for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$. □

Proof of lemma 3.

For a given $\eta^2 > 0$, write

$$
\begin{aligned}
|q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})| &= | - \eta^{-2}\log[\eta^{-2}/(\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))] - y_{it}\log[\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})/(\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))]| \\
&\leq |\eta^{-2}\log\eta^{-2}| + |(\eta^{-2} + y_{it})\log(\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))| + |y_{it}m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| \\
&\leq |\eta^{-2}\log\eta^{-2}| + |(\eta^{-2} + y_{it})(\eta^{-2} - 1 + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))| + |y_{it}m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| \\
&\leq |\eta^{-2}\log\eta^{-2}| + |(\eta^{-2} + y_{it})(\eta^{-2} - 1)| + |(\eta^{-2} + y_{it})\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})| + |y_{it}m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| \\
&\equiv q_1 + q_2 + q_3 + q_4
\end{aligned}
$$

The first and third inequalities follow from the triangle inequality, and the second follows from the inequality on the natural logarithm $\log u \leq u - 1$. Under assumption 3(b), the conditional mean is bounded, i.e., $E(y_{it}|\cdot) = \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) \leq \exp(|m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}|) \leq \sqrt{\kappa_y} < \infty$ where $E(y_{it}|\cdot) \equiv E(y_{it}|\mathbf{x}, \mathbf{z}, \boldsymbol{D}) = E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t)$.

The following bounds obtain on the terms $q_1,\ q_2,\ q_3$, and q_4.

$q_1 \equiv |\eta^{-2}\log\eta^{-2}| < \infty$ because η^2 is finite.

$E(q_2) \equiv |(\eta^{-2}+y_{it})(\eta^{-2}-1)| = |(\eta^{-2}-1)|E(\eta^{-2}+y_{it}) = |(\eta^{-2}-1)|(\eta^{-2}+E(y_{it})) = |(\eta^{-2}-1)|(\eta^{-2}+E[E(y_{it}|\cdot)]) = |(\eta^{-2}-1)|(\eta^{-2}+E\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})) \leq |(\eta^{-2}-1)|(\eta^{-2}+\sqrt{\kappa_y}) < \infty$. The first equality follows because $y_{it} \geq 0$ and $|(\eta^{-2}+y_{it})| = (\eta^{-2}+y_{it})$ and the conclusion follows from assumption 3(b).

$E(q_3|\cdot) = E((|(\eta^{-2}+y_{it})\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})|)\cdot) \leq \eta^{-2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})E(y_{it}|\cdot) \leq \eta^{-2}\sqrt{\kappa_y}+\kappa_y < \infty$ by assumption 3(b). By the law of iterated expectations (LIE), $E(q_3) = E(E(q_3|\cdot)) \leq \eta^{-2}\sqrt{\kappa_y}+\kappa_y$

$E(q_4|\cdot) = E(y_{it}m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}|\cdot) = m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}E(y_{it}|\cdot) \leq 1/2\sqrt{\kappa_y}\log\kappa_y < \infty$ by assumption 3(b). By the LIE, $E(q_4) = E(E(q_4|\cdot)) \leq 1/2\sqrt{\kappa_y}\log\kappa_y < \infty$

Combining terms shows that $E|q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})| < \infty$ from which it follows that $E|\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})| \leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E|q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})| \leq |\eta^{-2}\log\eta^{-2}| + |(\eta^{-2}-1)|(\eta^{-2}+\sqrt{\kappa_y}) + \eta^{-2}\sqrt{\kappa_y} + \kappa_y + 1/2\sqrt{\kappa_y}\log\kappa_y < \infty$ □

Proof of lemma 4. $\boldsymbol{w}(\boldsymbol{\delta})$ is non-singular (assumption 2(b)) which implies that $\boldsymbol{w}(\boldsymbol{\delta}) \neq \boldsymbol{w}(\boldsymbol{\delta}_o)$ if $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$ a.s. Since $\mathbf{x} \neq \mathbf{0}$ (assumption 2(a)), $\boldsymbol{m}(\boldsymbol{\delta}) \neq \boldsymbol{m}(\boldsymbol{\delta}_o)$ if $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$. By lemma 2, $\boldsymbol{m}(\boldsymbol{\delta})$ is full rank which implies $\boldsymbol{m}(\boldsymbol{\delta})\boldsymbol{\beta} \neq \boldsymbol{m}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o$ if $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$ and $\boldsymbol{\beta} \neq \boldsymbol{\beta}_o$.

Note that $\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}(\boldsymbol{\beta},\boldsymbol{\delta}) = \eta^{-2}\log\eta^{-2} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\big((\eta^{-2}+y_{it})\log(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})) - y_{it}m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}\big)$. Because the function $f(u) = \log(\eta^-2+\exp(u))$ is monotone in u, $f(\boldsymbol{m}(\boldsymbol{\delta})\boldsymbol{\beta}) \neq f(\boldsymbol{m}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$ if $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$ and $\boldsymbol{\beta} \neq \boldsymbol{\beta}_o$. From the forgoing, $\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta}) \neq \mathcal{Q}_n(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ if $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$ and $\boldsymbol{\beta} \neq \boldsymbol{\beta}_o$. This implies $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ is identified. □

Proof of theorem 1. Under the assumptions of lemma 2, $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ is identified. Also, under the assumptions of lemma 3, $E|\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})| < \infty$. The conclusion follows from Newey and McFadden (1994, lemma 2.2). □

Proof of proposition 1. Note that

$$
\begin{aligned}
&\lim_{N,T\to\infty}\frac{1}{NT}\Big|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{t=1}^{T}\sum_{\tau\neq t}\mathrm{cov}(y_{it},y_{j\tau}|\mathcal{F})\Big| = \lim_{N,T\to\infty}\frac{1}{NT}\Big|\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\tau\neq t}\mathrm{cov}(y_{it},y_{i\tau}|\mathcal{F})\Big| \\
&= \lim_{N,T\to\infty}\frac{2}{NT}\Big|\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\iota=1}^{T-t}\mathrm{cov}(y_{it},y_{it+\iota}|\mathcal{F})\Big| = \lim_{N,T\to\infty}\frac{2}{N}\Big|\sum_{i=1}^{N}\sum_{\iota=1}^{T-t}\Big(\frac{T-\iota}{T}\Big)\mathrm{cov}(y_{i1},y_{i1+\iota}|\mathcal{F})\Big| \\
&\leq \lim_{N\to\infty}\frac{2}{N}\sum_{i=1}^{N}\sum_{\iota=1}^{\infty}|\mathrm{cov}(y_{i1},y_{i1+\iota}|\mathcal{F})| \leq \lim_{N\to\infty}\frac{8\kappa_y}{N}\sum_{i=1}^{N}\sum_{\iota=1}^{\infty}\alpha_i^{\mathcal{F}}(\iota) \leq 8\bar{\alpha}\kappa_y
\end{aligned}
$$

The first equality follows from assumption 4(a) since for $i \neq j$, $\mathrm{cov}(y_{it}, y_{j\tau}|\mathcal{F}) = E(y_{it}y_{j\tau}|\mathcal{F}) - E(y_{it}|\mathcal{F})E(y_{j\tau}|\mathcal{F}) = E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t)E(y_{j\tau}|\mathbf{x}_\tau, \mathbf{z}_{j\tau}, \boldsymbol{D}_\tau) - E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t)E(y_{j\tau}|\mathbf{x}_\tau, \mathbf{z}_{j\tau}, \boldsymbol{D}_\tau) = 0$. The third equality uses conditional stationarity (assumption 4(c)). The conditional expectation $E(y_{it}|\mathcal{F})$ is bounded by $\sqrt{\kappa}$ for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$ thanks to assumption 3(b). The conclusion follows Rao (2009, theorem 9) and assumption 4(b). □

Proof of lemma 5. Proving uniform convergence in probability of $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})$ to $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})$ requires the verification of the following conditions.

First, the continuity of $q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta}) = -\eta^{-2}\log[\eta^{-2}/(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))] - y_{it}\log[\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})/(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))]$ at each $[\boldsymbol{\beta}', \boldsymbol{\delta}']' \in \boldsymbol{\Theta}$ holds by inspection.

Second, from lemma 3(see proof), $E|q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})| \leq |\eta^{-2}\log\eta^{-2}| + |(\eta^{-2}-1)|(\eta^{-2}+\sqrt{\kappa_y}) + \eta^{-2}\sqrt{\kappa_y} + \kappa_y + 1/2\sqrt{\kappa_y}\log\kappa_y < \infty$ a.s. for all $[\boldsymbol{\beta}', \boldsymbol{\delta}']' \in \boldsymbol{\Theta}$.

Coupled with assumptions 5 and that the data are random (i.e., defined on a probability space), the conclusion follows from Wooldridge (2010, theorem 12.1). □

Proof of theorem 2. The proof of consistency requires the verification of the following conditions.

First, under the assumptions of theorem 1, $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})$ is uniquely minimised at $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$.

Second, under the assumptions of lemma 5, $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})$ is continuous and $\sup_{\boldsymbol{\beta}\in\boldsymbol{B},\boldsymbol{\delta}\in\boldsymbol{\Delta}} |\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta}) - \mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})| \xrightarrow{p} 0$.

Coupled with assumption 5, the conclusion follows from Newey and McFadden (1994, theorem 2.1). $\square$

Proof of lemma 6. (a)

$$E[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] \equiv E\begin{bmatrix} \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}} \\ \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \end{bmatrix}\Bigg|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o \\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}$$

Taking expectations of partitions of $s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ separately,

$$E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o \\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}} = E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - y_{it})}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)} m_{it}(\boldsymbol{\delta}_o)'\Big]$$
$$= E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - E(y_{it}|\cdot))}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)} m_{it}(\boldsymbol{\delta}_o)'\Big] = \mathbf{0}$$

$$E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o \\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}} = E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) - y_{it})}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})} \frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o \\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}$$
$$= E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - E(y_{it}|\cdot))}{\eta^{-2} + \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)} \Big(\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}}\Big)\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o \\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}\Big] = \mathbf{0}$$

The second equalities for both partitions of $E[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$ follow from the law of iterated expectations, and that the conditional expectation $E(y_{it}|\cdot) \equiv \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$.

(b)

$$\mathrm{var}[(NT)^{1/2} s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] = (NT)E[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)']$$
$$= E\Big[\frac{1}{NT}\Big[\sum_{i=1}^N \sum_{t=1}^T s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)\Big]\Big[\sum_{i=1}^N \sum_{t=1}^T s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)\Big]'\Big]$$
$$= E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{j\neq i} \Big[\sum_{t=1}^T s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta})_o\Big]\Big[\sum_{t=1}^T s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)\Big]'\Big]$$
$$= E\Big[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T E[s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)'|\cdot] + \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \sum_{\tau\neq t} E[s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)s_{i\tau}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)'|\cdot]\Big]$$

The last line applies the law of iterated expectations and uses assumption 4(a), noting that

$E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{j\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot] = \mathbf{0}_{k_\theta\times k_\theta}$ for $i\neq j$ (see also proposition 1), i.e., considering the part dependent on $\{y_{it}\}_{i=1,t=1}^{N,T}$, $E[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{it})(\exp(m_{j\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{j\tau})|\cdot] = E(y_{it}y_{j\tau}|\cdot) - \exp((m_{it}(\boldsymbol{\delta}_o)+m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) = E(y_{it}|\cdot)E(y_{j\tau}|\cdot) - \exp((m_{it}(\boldsymbol{\delta}_o)+m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) = \exp((m_{it}(\boldsymbol{\delta}_o)+m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) - \exp((m_{it}(\boldsymbol{\delta}_o)+m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) = 0$ for $i\neq j$.

First, consider

$$E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot] = \frac{\eta^{-4}E[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{it})^2|\cdot]}{(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))^2}\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$$

where $\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) \equiv \begin{bmatrix} m_{it}(\boldsymbol{\delta}_o)'m_{it}(\boldsymbol{\delta}_o) & m_{it}(\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) \\ \mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'m_{it}(\boldsymbol{\delta}_o) & \mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\end{bmatrix}$ and $\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) \equiv \frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}$. The term $E[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{it})^2|\cdot] = E(y_{it}^2|\cdot) - \exp(2m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$. Using that $E(y_{it}^2|\cdot) = \mathrm{var}(y_{it}|\cdot)+(E(y_{it}|\cdot))^2$ and $\mathrm{var}(y_{it}|\cdot) = \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)+\eta^2\exp(2m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$ for the negative binomial model (assumption 1), $E[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{it})^2|\cdot] = \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)+\eta^2\exp(2m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$. Simplifying terms, we have $E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot] = \frac{1}{(\eta^{-2}+\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$.

Second, consider $E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{i\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot] = \frac{\eta^{-4}E[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{it})(\exp(m_{i\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{i\tau})|\cdot]}{(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))(\eta^{-2}+\exp(m_{i\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ where $\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) \equiv \begin{bmatrix} m_{it}(\boldsymbol{\delta}_o)'m_{i\tau}(\boldsymbol{\delta}_o) & m_{it}(\boldsymbol{\delta}_o)'\mu_{i\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) \\ \mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'m_{i\tau}(\boldsymbol{\delta}_o) & \mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'\mu_{i\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\end{bmatrix}$

The term $E[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{it})(\exp(m_{i\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)-y_{i\tau})|\cdot] = E(y_{it}y_{i\tau}|\cdot) - \exp((m_{it}(\boldsymbol{\delta}_o)+m_{i\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o)$. Using that $\mathrm{cov}(y_{it},y_{i\tau}|\cdot) = E(y_{it}y_{i\tau}|\cdot) - E(y_{it}|\cdot)E(y_{i\tau}|\cdot)$,

$$E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{i\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot] = \frac{\eta^{-4}\mathrm{cov}(y_{it},y_{i\tau}|\cdot)}{(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))(\eta^{-2}+\exp(m_{i\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$$

Let us apply bounds on terms. $||\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 = ||\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 \leq ||m_{it}(\boldsymbol{\delta}_o)'m_{i\tau}(\boldsymbol{\delta}_o)||_2 + 2||m_{it}(\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 \leq ||m_{it}(\boldsymbol{\delta}_o)||_2^2 + 2||m_{it}(\boldsymbol{\delta}_o)||_2\cdot||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2 = (||m_{it}(\boldsymbol{\delta}_o)||_2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2)^2 \leq (\kappa_m + 2\kappa_d\kappa_w\kappa_z)^2$. For the first part, the following bound holds.

$$\Big|\Big|E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot]\Big]\Big|\Big|_2 = \Big|\Big|E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{(\eta^{-2}+\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\Big]\Big|\Big|_2$$
$$\leq \eta^2(\kappa_m+2\kappa_d\kappa_w\kappa_z)^2$$

The inequality follows by applying the bound on $\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ and noting that $(\eta^{-2}+\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)) > \eta^{-2}$. For the second part,

$$\Big|\Big|E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\tau\neq t}E[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{i\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot]\Big]\Big|\Big|_2$$
$$= \Big|\Big|E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\tau\neq t}\frac{\eta^{-4}\text{cov}(y_{it},y_{i\tau}|\cdot)}{(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))(\eta^{-2}+\exp(m_{i\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\Big]\Big|\Big|_2$$
$$= \Big|\Big|E\Big[\frac{2}{N}\sum_{i=1}^{N}\sum_{\iota=1}^{T-1}\Big(\frac{T-\iota}{T}\Big)\Big(\frac{\eta^{-4}\text{cov}(y_{i1},y_{i1+\iota}|\cdot)}{(\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))(\eta^{-2}+\exp(m_{i\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\Big)\Big]\Big|\Big|_2$$
$$\leq E\Big[\frac{2}{N}\sum_{i=1}^{N}\sum_{\iota=1}^{T-1}|\text{cov}(y_{i1},y_{i\iota+1}|\cdot)|\cdot||\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2\Big]$$
$$\leq 2\eta^2(\kappa_m+2\kappa_d\kappa_w\kappa_z)^2E\Big[\frac{1}{N}\sum_{i=1}^{N}\sum_{\iota=1}^{T-1}|\text{cov}(y_{i1},y_{i\iota+1}|\cdot)|\Big] \leq 8\bar{\alpha}\kappa_y\eta^2(\kappa_m+2\kappa_d\kappa_w\kappa_z)^2$$

The second equality follows assumption 4(c) and the conclusion uses the result in proposition 1. Both parts are bounded, and the proof is complete. □

Proof of lemma 7 . We establish the following bounds on the partitions of $H_n(\boldsymbol{\beta},\boldsymbol{\delta})$ for any $\boldsymbol{\beta} \in \boldsymbol{B}$ and $\boldsymbol{\delta} \in \boldsymbol{\Delta}$.

$$
\begin{aligned}
E\Big[\Big|\Big|\frac{\partial \mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\Big|\Big|_2\Big] &= E\Big[\Big|\Big|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\Big|\Big|_2\Big] \\
&= E\Big[\Big|\Big|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}m_{it}(\boldsymbol{\delta})'m_{it}(\boldsymbol{\delta})\Big|\Big|_2\Big] \\
&= E\Big[\Big|\Big|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(\eta^{-2}+y_{it})}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}m_{it}(\boldsymbol{\delta})'m_{it}(\boldsymbol{\delta})\Big|\Big|_2\Big] \\
&\leq E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(\eta^{-2}+E(y_{it}|\cdot))}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}||m_{it}(\boldsymbol{\delta})'m_{it}(\boldsymbol{\delta})||_2\Big] \\
&\leq E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(\eta^{-2}+\sqrt{\kappa_y})}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}||m_{it}(\boldsymbol{\delta})'m_{it}(\boldsymbol{\delta})||_2\Big] \\
&\leq E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}||m_{it}(\boldsymbol{\delta})'m_{it}(\boldsymbol{\delta})||\Big] = E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{2NT}||\boldsymbol{m}(\boldsymbol{\delta})||_2^2\Big] \\
&\leq \frac{\kappa_m^2}{2}(\eta^{-2}+\sqrt{\kappa_y}) < \infty
\end{aligned}
$$

$$
\begin{aligned}
&E\Big[\Big|\Big|\frac{\partial \mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] = E\Big[\Big|\Big|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&\leq E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}m_{it}(\boldsymbol{\delta})'\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'} \\
&+\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-y_{it})}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&\leq E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}m_{it}(\boldsymbol{\delta})'\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&+E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-y_{it})}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{(\eta^{-2}+y_{it})}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}m_{it}(\boldsymbol{\delta})'\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&+E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\Big(\frac{\eta^{-2}}{1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}-\frac{y_{it}}{1+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\Big)\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(\eta^{-2}+E(y_{it}|\cdot))}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}\Big|\Big|m_{it}(\boldsymbol{\delta})'\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&+E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\Big(\frac{\eta^{-2}}{1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}-\frac{E(y_{it}|\cdot)}{1+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\Big)\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&\leq E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|m_{it}(\boldsymbol{\delta})'\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&+E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&\leq E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|m_{it}(\boldsymbol{\delta})\Big|\Big|_2\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&+E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}'}\Big|\Big|_2\Big] \\
&\leq \kappa_d\kappa_w\kappa_z(\eta^{-2}+\sqrt{\kappa_y})(1/2\log\kappa_y+2)<\infty
\end{aligned}
$$

The conclusion uses bounds on derivative terms $\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}$ and $\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}$ established in lemma 9.

$$
\begin{aligned}
&E\Big[\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big] = E\Big[\Big|\Big|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&\le E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\\
&+\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-y_{it})}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&\le E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&+E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-y_{it})}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&=E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{(\eta^{-2}+y_{it})}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&+E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\Big(\frac{\eta^{-2}}{1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}-\frac{y_{it}}{1+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\Big)\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&=E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{(\eta^{-2}+E(y_{it}|\cdot))}{(2+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&+E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\Big(\frac{\eta^{-2}}{1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}-\frac{E(y_{it}|\cdot)}{1+\eta^{2}\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\Big)\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&\le E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]+E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&\le E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\Big|\Big|_2^2\Big]+E\Big[\frac{(\eta^{-2}+\sqrt{\kappa_y})}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2\Big]\\
&\le \kappa_\beta\kappa_d\kappa_w\kappa_z(\eta^{-2}+\sqrt{\kappa_y})(1+5\kappa_d)
\end{aligned}
$$

The conclusion uses bounds on derivative terms $\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}$ and $\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}$ established in lemma 9.

Combining terms shows that $E[||H_n(\boldsymbol{\beta},\boldsymbol{\delta})||_2] < \infty$ for all $\boldsymbol{\beta} \in \boldsymbol{B}$ and $\boldsymbol{\delta} \in \boldsymbol{\Delta}$, and the proof is complete. $\square$

Proof of lemma 8. The expectation of the hessian has the following partitions: Ξ

$$
\begin{aligned}
E\Big[\frac{\partial \mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}} &= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}\\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))^2}m_{it}(\boldsymbol{\delta}_o)'m_{it}(\boldsymbol{\delta}_o)\Big]\\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}m_{it}(\boldsymbol{\delta}_o)'m_{it}(\boldsymbol{\delta}_o)\Big]\\
&= E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a]
\end{aligned}
$$

where $\boldsymbol{m}_a \equiv \boldsymbol{D}_a(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)^{1/2}\boldsymbol{m}(\boldsymbol{\delta}_o)$ and $\boldsymbol{D}_a(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ is an $NT\times NT$ diagonal matrix whose $((t-1)N+i)$'th diagonal element is $\frac{\eta^{-2}}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}$.

$$
\begin{aligned}
&E\Big[\frac{\partial \mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}} = E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}\\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}m_{it}(\boldsymbol{\delta})'\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\\
&+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-y_{it})}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}\\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}m_{it}(\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\Big]\\
&= E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{\mu}_a]
\end{aligned}
$$

where $\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) \equiv \frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}$, $\boldsymbol{\mu}_a \equiv \boldsymbol{D}_a(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)^{1/2}\boldsymbol{\mu}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$, and $\boldsymbol{\mu}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ is an $NT\times k_\delta$ matrix whose $((t-1)N+i)$'th row is $\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$.

$$
\begin{aligned}
&E\Big[\frac{\partial \mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}} = E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}\\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\eta^{-2}+y_{it})\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}))^2}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}\\
&+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-y_{it})}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}}\\
&= E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\Big]\\
&= E[\frac{1}{NT}\boldsymbol{\mu}_a'\boldsymbol{\mu}_a]
\end{aligned}
$$

Combining terms, the expectation of the hessian matrix is $E[H_n(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)] = E\frac{1}{NT}\begin{bmatrix}\boldsymbol{m}_a'\boldsymbol{m}_a & \boldsymbol{m}_a'\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_a'\boldsymbol{m}_a & \boldsymbol{\mu}_a'\boldsymbol{\mu}_a\end{bmatrix}$

To prove positive definiteness of $E[H_n(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)]$, first note that $E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a]$ is invertible using that $E[\frac{1}{NT}\boldsymbol{m}'\boldsymbol{m}]$ is positive definite (lemma 2) and each diagonal element of $\boldsymbol{D}_a(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ is positive and finite.

Second, the Schur complement of $E[H_n(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)]$ given by

$$
\begin{aligned}
&E[\frac{1}{NT}\boldsymbol{\mu}_a'\boldsymbol{\mu}_a] - E[\frac{1}{NT}\boldsymbol{\mu}_a'\boldsymbol{m}_a](E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{\mu}_a]\\
&= E[\hat{\boldsymbol{\mu}}_a'\hat{\boldsymbol{\mu}}_a] \text{ is positive definite.}
\end{aligned}
$$

where $\hat{\boldsymbol{\mu}}_a = \frac{1}{NT}\boldsymbol{\mu}_a'(\mathbf{I}_{NT} - \boldsymbol{m}_a(E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}\boldsymbol{m}_a')$ and $\mathbf{I}_{NT}$ denotes an $NT\times NT$ identity matrix.

$\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$ can be expressed as $\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o) = -\rho_{o,2}\sum_{j\neq i} w_{ijt}(\boldsymbol{\delta}_o)(\sum_{\substack{l\neq i\\ l\neq j}} x_{lt})\boldsymbol{d}_{ijt}'$ where $\rho_{o,2}$ is the third element in $\boldsymbol{\beta}_o$. We show that $(\sum_{j\neq i}(\rho_{o,2}w_{ijt}(\boldsymbol{\delta}_o)(\sum_{\substack{l\neq i\\ l\neq j}} x_{lt}))^2)^{1/2} < \infty$ a.s. $|\rho_{o,2}| \leq \kappa_\beta < \infty$ by assumption 5. $w_{ijt}(\boldsymbol{\delta}_o) \leq \frac{\kappa_w}{N}$ a.s. for all $i,j = 1,\ldots,N$, $j\neq i$, by lemma 1. $|\sum_{\substack{l\neq i\\ l\neq j}} x_{lt}| \leq \sum_{\substack{l\neq i\\ l\neq j}} |x_{lt}| \leq (N-2)\kappa_z$ a.s. for all $t = 1,\ldots,T$. Combining terms, $(\sum_{j\neq i}(\rho_{o,2}w_{ijt}(\boldsymbol{\delta}_o)(\sum_{\substack{l\neq i\\ l\neq j}} x_{lt}))^2)^{1/2} \leq \kappa_\beta\kappa_w\kappa_z < \infty$. Also, $(\sum_{j\neq i}(\rho_{o,2}w_{ijt}(\boldsymbol{\delta}_o)(\sum_{\substack{l\neq i\\ l\neq j}} x_{lt}))^2)^{1/2} >$

0 since $\rho_{o,2} \neq 0$ (assumption 6(b)) and $\mathbf{x} \neq \mathbf{0}$ a.s. assumption 2(a). Because $\boldsymbol{D}_a(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ is positive definite, $\boldsymbol{\mu}_a \equiv \boldsymbol{D}_a(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)^{1/2}\boldsymbol{\mu}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ is full rank under assumption 6(a), i.e., $\text{rank}(\boldsymbol{\mu}_a) = k_\delta$. Note that $\boldsymbol{m}_a(E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}\boldsymbol{m}_a'$ is a projection matrix. Because the rank of $\boldsymbol{m}_a$ is k_β (using lemma 2 and that $\boldsymbol{D}_a(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ is positive definite), the rank of the projection matrix is k_β while the rank of $\mathbf{I}_{NT} - \boldsymbol{m}_a(E[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}\boldsymbol{m}_a'$ is $NT - k_\beta$. $NT - k_\beta - k_\delta > 0$ for any estimable model hence it follows that the $\text{rank}(\hat{\boldsymbol{\mu}}_a) = k_\delta$ and the Schur complement $E[\hat{\boldsymbol{\mu}}_a'\hat{\boldsymbol{\mu}}_a]$ is positive definite.

The conclusion follows using the result in Boyd and Vandenberghe (2004, sect A.5.5) on the positive definiteness of symmetric matrices. □

Proof of theorem 3. Under the assumptions of theorem 2, $\hat{\boldsymbol{\theta}} \equiv [\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}']'$ is consistent. Under the assumptions of lemma 6, $E[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] = \mathbf{0}$ and each element of the second moment of $(NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ is finite. Under the assumptions of lemma 7, the hessian matrix $H_n(\boldsymbol{\beta}, \boldsymbol{\delta})$ is bounded in absolute value for all $[\boldsymbol{\beta}', \boldsymbol{\delta}']' \in \boldsymbol{\Theta}$. Under the assumptions of lemma 8, $E[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$ is positive definite. The continuous differentiability of the score function on the interior of $\boldsymbol{\Theta}$ holds by inspection. In addition to assumption 7, $\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_\theta)$ by Wooldridge (2010, theorem 12.3) where $\mathbf{V}_\theta \equiv \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$, $\mathbf{A}_o \equiv E[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$, and $\mathbf{B}_o \equiv \text{var}((NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o))$. □

Proof of corollary 1. $\sqrt{NT}(\varrho_n(\hat{\boldsymbol{\theta}}) - \varrho(\boldsymbol{\theta}_o)) = \sqrt{NT}(\varrho_n(\hat{\boldsymbol{\theta}}) - \varrho(\hat{\boldsymbol{\theta}})) + \sqrt{NT}(\varrho(\hat{\boldsymbol{\theta}}) - \varrho(\boldsymbol{\theta}_o))$. The first term converges to zero in probability. The continuous differentiability of $\varrho(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ holds by inspection. By the mean-value theorem, $\sqrt{NT}(\varrho(\hat{\boldsymbol{\theta}}) - \varrho(\boldsymbol{\theta}_o)) = \varrho'(\bar{\boldsymbol{\theta}})\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ where $\bar{\boldsymbol{\theta}}$ is a mean value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$. Because $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$ and $\bar{\boldsymbol{\theta}}$ is trapped $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$, $\varrho'(\bar{\boldsymbol{\theta}}) \xrightarrow{p} \varrho'(\boldsymbol{\theta}_o)$. It follows that $\sqrt{NT}(\varrho(\hat{\boldsymbol{\theta}}) - \varrho(\boldsymbol{\theta}_o)) = \varrho'(\boldsymbol{\theta}_o)\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \varrho'(\boldsymbol{\theta}_o)'\mathbf{V}_\theta\varrho'(\boldsymbol{\theta}_o))$. □