R&D spillover effects on firm innovation - Estimating the spatial matrix from panel data *

Emmanuel S. Tsyawo†

Latest version here

November 2019

Abstract

Firms' research and development (R&D) efforts are known to have spillover effects on other firms' outcomes, e.g., innovation. Quantifying R&D spillover effects requires a spatial matrix that characterises the interaction structure between firms. Estimates can be biased if the spatial matrix is misspecified. This paper proposes a parsimonious approach to estimating the spatial matrix and parameters from panel data, when the spatial matrix is partly or fully unknown. The approach is robust to misspecification of the spatial matrix, and it is applicable to single index models. It allows asymmetry and time-variation in the spatial matrix. The paper establishes consistency and asymptotic normality of the MLE under conditional independence and conditional strong-mixing assumptions on the outcome variable. On firm innovation, I find positive spillover and private effects of R&D. I 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, 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 matrix, negative binomial model

JEL classification: C21, L25

*I thank 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 outcomes, e.g., 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 largely rely on pre-specified spatial matrices. A spatial matrix[1] defines the strength of connectivity between firms, and conclusions differ by the choice of spatial matrix. Knowledge spillovers have policy implications and accurately quantifying them aids to fine-tune policy,[2] 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 typically obtained by weighting the R&D of other firms using a spatial matrix. In the literature, the spatial matrix is often assumed known (e.g., Bloom, Schankerman, and Van Reenen (2013) and König, Liu, and Zenou (2019)). Misspecifying the spatial matrix (and spillover measure by extension) leads to misleading conclusions and policy recommendations (see e.g., Bhattacharjee and Jensen-Butler (2013)). In light of the foregoing challenge, this paper introduces a flexible approach that is robust to misspecification of the spatial matrix, models and estimates the spatial matrix from panel data, and quantifies the spillover effect of R&D on innovation.

The proposed method in this paper does offer a number of advantages. It models the

[1] In the duopoly case, the spatial matrix is a 2×2 matrix where the magnitude of off-diagonal elements determines the extent and magnitude of spillover effects.

[2] Spillovers (positive externalities), from a social perspective, imply under-investment in R&D 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) finds limited knowledge spillovers of small firms and recommends a reconsideration of higher R&D tax credits given to small firms.

structure of interactions between firms across time using spatial covariates (indicators of forms of proximity between firms, e.g., geographic and product market proximity) and a parameter vector of finite length. Directly estimating a spatial matrix from data typically involves a parameter vector that is quadratically increasing in the number of firms (e.g., Manresa (2013) and Soale and Tsyawo (2019)). This paper reduces the parameter space associated with the spatial matrix to one of finite length. The proposed approach is applicable to the class of single-index models[3], e.g., negative binomial, logit, linear, and quantile regression. The wide applicability of the approach is particularly useful for quantifying spillovers and spatial dependence because some firm outcomes, e.g., innovation, are measured using non-continuous variables whence the need for non-linear models. The approach is parsimonious, data-driven, and testable. It allows a time-varying and asymmetric spatial matrix that admits several spatial covariates simultaneously. The relevance of different forms of proximity between firms (measured by the respective spatial covariates) is statistically assessable, and it sheds light on the channels of transmission of spillovers between firms. Data-driven determination of relevant spatial covariates obviates the arbitrary choice of spatial matrix, thus rendering the proposed approach robust to misspecification of the spatial matrix unlike most of the extant literature. The proposed method is easy to implement and tools of M-estimation and inference are applicable. I contribute to the literature in two ways. First, I quantify spillover effects when the interaction structure is partly or fully unknown. Second, the econometric approach is applicable to the class of single-index models.

The paper applies the approach to the negative binomial model because the outcome variable, citation-weighted patent counts, is discrete. I provide results on consistency and asymptotic normality of the maximum likelihood estimator under conditional stationarity (across time), conditional independence (across firms), and strong-mixing (over time) assumptions on the outcome. Since the econometric framework only requires conditional independence, the outcome across firms can be (unconditionally) dependent, and the outcome

[3]Though the approach is applicable to the class of all single-index models, I limit the scope to single-index models with known link functions for simplicity.

of firms across time can be sequentially dependent.

The 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)), I use variables including citation-weighted patent counts (as outcome variable for innovation) and R&D stock (for R&D) in estimation. The empirical results show sensitivity of spillover effects (and other parameters) to the choice of spatial matrix. Using a robust specification that controls for several forms of proximity between firms, I 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. The empirical results confirm the relevance of geographic proximity and product market proximity. The relevance of spatial covariates associated with past R&D and patenting behaviour of firms are also established. 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 estimated interaction structure between firms is asymmetric and time-varying.

The rest of the paper is organised as follows. Section 2 provides a review of the literature and section 3 presents the approach with the identification result. Section 4 provides empirical and econometric extensions to the proposed approach. Section 5 covers the estimation algorithm and section 6 presents asymptotic results of the MLE estimator. 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; I identify four 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)).[4] A second approach employs industry distance to the frontier[5] 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, 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.

At the econometric level, much progress is realised in the last two decades with respect to estimating spillover effects and spatial dependence in general. A first category comprises spatial econometric models, e.g., the Spatial Autoregressive (SAR) model that spatially lags the outcome variable (e.g., Anselin, Le Gallo, and Jayet (2008)). SAR models typically rely on pre-specified spatial matrices. A second category directly estimates the spatial matrix as a set of parameters. This 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 category models individual elements of the spatial matrix up to a finite number of parameters and estimates all parameters using panel data (Kapetanios, Mitchell, and Shin, 2014; Pinkse, Slade, and Brett, 2002). The proposed approach falls in this last category. Specifically, this paper's approach extends to both linear and non-linear single-index models. Also, in employing several spatial covariates, this paper simultaneously controls for several forms of proximity between firms whence the approach's robustness to misspecification of the spatial matrix.

As regards constructing spillover measures and quantifying R&D spillovers, one notes a number of shortcomings of existing approaches. Exogenously constructing spillover measures

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

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

from pre-specified metrics can lead to misleading results if the spatial matrix is misspecified. 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 (i.e., relevant forms of proximity between firms), 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), or product market (Bloom, Schankerman, and Van Reenen, 2013), knowledge spillovers can be explained by other forms of proximity that are often overlooked in the literature, e.g., social networks and labour mobility.[6]

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 (2019). These approaches, however, suppose foreknowledge of the spatial matrix. 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 measure of 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 models for estimation.

3 The model

Modelling spatial dependence using the proposed approach spans the class of single-index models, e.g., linear, quantile, logit, and probit regressions. In this paper, I focus on

[6]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, e.g., geographic proximity.

the negative binomial model as the outcome variable (citation-weighted patent counts) is discrete. Let data $\{y_{it}, x_{it}, \mathbf{z}_{it}, \{\boldsymbol{d}_{ijt}\}_{j=1}^{N}\}_{i=1,t=1}^{N,T}$, be defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$. The outcome y_{it}, covariates $\mathbf{z}_{it}$ and $\mathbf{x}_t = [x_{1t}, \ldots, x_{Nt}]'$ (where $\mathbf{x}_t$ is R&D), and spatial covariates $\boldsymbol{d}_{ijt}$ (e.g. geographic distance) are observed. Let $\boldsymbol{D}_t \equiv \{\boldsymbol{d}_{ijt} : i, j = 1, \ldots, N,\ j \neq i\}$ and $\boldsymbol{D} \equiv \bigcup_{t=1}^{T} \boldsymbol{D}_t$. The conditional distribution assumption is stated in the following.

Assumption 1. *Conditional on $\mathbf{x}_t, \mathbf{z}_{it}$, and $\boldsymbol{D}_t$, y_{it} follows the negative binomial distribution with a finite dispersion parameter $\eta^2 > 0$.*[7]

3.1 The conditional expectation

Under assumption 1, the conditional expectation and conditional variance, respectively are specified as follows

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

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

where $w_{ijt}(\boldsymbol{\delta}) \equiv \dfrac{\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\sum_{j \neq i}^{N} \exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}$, $\forall\, i = 1, \ldots, N,\ j \neq i,\ t = 1, \ldots, T$. The function $\exp(\cdot)$ in $w_{ijt}(\cdot)$ 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 $\boldsymbol{\beta} \equiv [\rho_0, \rho_1, \rho_2, \boldsymbol{\gamma}']' \in \boldsymbol{B}$, $\boldsymbol{\delta} \in \boldsymbol{\Delta}$, $\boldsymbol{\Theta} \equiv \boldsymbol{B} \times \boldsymbol{\Delta}$, 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.

[7] For ease of exposition, η^2 is treated as known.

Spatial covariates, denoted by the $k_\delta \times 1$ vector $\boldsymbol{d}_{ijt}$, measure different forms of proximity 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 past 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 between i and j, $j \neq i$ at time t 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 this paper's approach is applicable; other models 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 $\mathbf{x}_t, \mathbf{z}_{it}$, and $\boldsymbol{D}_t$

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

where $g : \mathbb{R} \to \mathbb{R}$ is a link function, and $\nu(\cdot|\cdot)$ denotes the conditional functional, namely, conditional expectation (e.g., linear regression and the negative binomial model eq. (3.1)), the conditional quantile (quantile regression), and 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 the design matrix $\boldsymbol{m}(\boldsymbol{\delta})$. $q_{it}(\boldsymbol{\beta}, \boldsymbol{\delta})$ of other single index 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 whose corresponding coefficient in $\boldsymbol{\delta}$ is non-zero, $\boldsymbol{w}_t(\boldsymbol{\delta})$ varies by time $t = 1, \ldots, T$, and the structure of interactions need not be stable through time. Also, I do not assume $\boldsymbol{d}_{ijt} = \boldsymbol{d}_{jit},\ j \neq i$ whence possible asymmetry in $\boldsymbol{w}(\boldsymbol{\delta})$. Properties of the proposed approach like time-variation and asymmetry can be assessed statistically by testing the significance of elements in $\boldsymbol{\delta}$ corresponding to time-varying and/or asymmetric elements in $\boldsymbol{D}$.

3.3 Identification

Identification of maximum likelihood estimators is important for consistency of the MLE. The case of this paper does differ from standard econometric cases because the spatial matrix in addition to other parameters $\boldsymbol{\beta}$ ought to be estimated. This subsection addresses identification. The following assumptions are useful in bounding the design matrix $\boldsymbol{m}(\boldsymbol{\delta})$ for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$.

Assumption 2. *(a) There exists a positive constant* κ_w, $0 < \kappa_w^{8.5} < \infty$ *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, $1 \leq \kappa_y^{8.5} < \infty$, *such that* $\sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} ||m_{it}(\boldsymbol{\delta})||_2 \leq \kappa_m$, $\sup_{\boldsymbol{\beta} \in \boldsymbol{B}} \kappa_m ||\boldsymbol{\beta}||_2 \leq 1/2 \log \kappa_y$ *for each* $i = 1, \ldots, N$, $t = 1, \ldots, T$, *and* $||\boldsymbol{\xi}||_p$ *denotes* $p-$*norm applied to vectorised* $\boldsymbol{\xi}$, $p \geq 1$. *(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 *and* κ_z *such that* $\max_{\substack{1 \leq j \leq N \\ 1 \leq t \leq T \\ j \neq i}} ||\boldsymbol{d}_{ijt}||_2 \leq \kappa_d$, *and each element in* $\boldsymbol{z}$ *is bounded in absolute value by* κ_z *a.s.,* $0 < \kappa_d^{8.5} < \infty$, $0 < \kappa_z^{8.5} < \infty$.

Note that $\sup_{\boldsymbol{\beta}\in\boldsymbol{B},\boldsymbol{\delta}\in\boldsymbol{\Delta}} |m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| \leq 1/2 \log \kappa_y$ follows from assumption 2(b) because $|m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}| = ||m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}||_2 \leq ||m_{it}(\boldsymbol{\delta})||_2||\boldsymbol{\beta}||_2 \leq 1/2 \log \kappa_y$. I introduce assumption 2(a & b) to ensure column sum boundedness of $\boldsymbol{w}(\boldsymbol{\delta})$. Assumption 2(c) is standard, and it is needed to ensure the boundedness of the objective function $\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})$. By construction, the row sums of $\boldsymbol{w}(\boldsymbol{\delta})$ are bounded at 1. The following lemma bounds the column sums of $\boldsymbol{w}(\boldsymbol{\delta})$.

Lemma 1. *Under assumption 2(a),* $\sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \max_{\substack{1\leq j\leq N \\ 1\leq t\leq T}} w_{ijt}(\boldsymbol{\delta}) \leq \frac{\kappa_w}{N}$*, for any* $i \neq j$*, and the column sums of* $\boldsymbol{w}(\boldsymbol{\delta})$ *are bounded in absolute value for all* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$*.*

Full column rank condition of the design matrix is useful for identification and consistency of the estimator. Establishing the full rank condition of the design matrix requires the following assumptions.

Assumption 3. *(a)* $\boldsymbol{z} \equiv [\mathbf{1}_{NT}, \mathbf{x}, \mathbf{z}]$ *is full column rank. (b)* $\boldsymbol{w}_t(\boldsymbol{\delta})$ *is non-singular,* $t = 1, .., T$ *for all* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$*. (c) All diagonal elements of* $\boldsymbol{w}_t(\boldsymbol{\delta})$ *are zero,* $t = 1, .., T$ *for all* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$*.*

Assumption 3(a) is a standard identification assumption. Assumption 3(b) only fails in pathological cases; it holds if there is enough variation in $\boldsymbol{D}_t$ for $t = 1, \ldots, T$. Note that $\mathbf{x} \neq \mathbf{0}$ a.s. follows from assumption 3(a), and it is necessary (coupled with assumption 3(b)) for $\boldsymbol{w}\mathbf{x}$ to possess independent variation from $\mathbf{x}$ in $\boldsymbol{z}$. Assumption 3(c) is standard in the spatial econometrics literature (e.g., Anselin, Le Gallo, and Jayet (2008)); it disallows a firm from generating spillovers on itself.

The following lemma shows that the full column rank condition of the design matrix $\boldsymbol{m}$ holds for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ under assumption 3(a-c).

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

The following lemma establishes a bound on $E|\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})|$.

Lemma 3. *Under assumptions 1 and 2(b),* $E|\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})| < \infty$ *for all* $\boldsymbol{\beta} \in \boldsymbol{B}$ *and* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$*.*

The following lemma provides the identification result.

Lemma 4 (Identification)**.** *Under assumptions 2(b), and 3(a-c),* $\boldsymbol{\theta}_o \equiv [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ *is identified where* $\boldsymbol{\theta}_o$ *denotes the true parameter.*

The identification result (lemma 4) relies on lemma 10 provided in appendix A. The uniqueness of the true parameter vector $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ is established thanks to lemmas 1 to 4. The result is provided in the following theorem.

Theorem 1. *Under assumptions 2(a&b), and 3(a-c),* $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta}) \equiv \mathrm{E}[\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})]$ *has a unique minimum at* $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$*.*

3.4 Parameters of interest

I explore parameters that are interesting from a policy perspective. Though these parameters can have varied interpretation by application (see section 4.2 for examples), I interpret them in the context of R&D spillover effects on innovation using the negative binomial set-up in eq. (3.1). 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. For a specific firm i, the private effect of its R&D at time t is $\mathcal{PE}_{it} \equiv \frac{\partial E(y_{it}|\cdot)}{\partial x_{it}} = \rho_1 E(y_{it}|\cdot)$, and averaged across time, $\mathcal{PE}_i = T^{-1}\sum_{t=1}^{T}\mathcal{PE}_{it} = \rho_1 T^{-1}\sum_{t=1}^{T}E(y_{it}|\cdot)$, where $E(y_{it}|\cdot) \equiv E(y_{it}|\mathbf{x}_t, \mathbf{z}_{it}, \boldsymbol{D}_t)$ (eq. (3.1)). The average private effect across all firms and time is given by $\mathcal{PE} = N^{-1}\sum_{i=1}^{N}\mathcal{PE}_i$. The conditional expectation (eq. (3.1)) does not allow heterogeneity in ρ_1 across firms unlike, Manresa (2013) and Soale and Tsyawo (2019, sect. 6). It can, however, allow variation in the private effect through time or by firm.

Spillover effect: The spillover effect measures the impact of firm j's R&D on another i's innovation, $i \neq j$. 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 firm j's R&D on firm i's, $j \neq i$, innovation 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 generated 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 $t = 1, \ldots, T$.

Social effect: The social effect of R&D on innovation sums private and spillover effects. It denotes the effect on aggregate innovation of the R&D efforts of all firms. 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 $\mathcal{SE}_{i.t} = \sum_{j=1}^{N} \frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \frac{\partial E(y_{it}|\cdot)}{\partial x_{it}} + \sum_{j\neq i} \frac{\partial E(y_{it}|\cdot)}{\partial x_{jt}} = \mathcal{PE}_{it} + \sum_{j\neq i} \mathcal{S}p\mathcal{E}_{ijt} = \mathcal{PE}_{it} + \mathcal{S}p\mathcal{E}_{i.t}$. Average social effects across firms and time obtains as $\mathcal{SE} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathcal{SE}_{i.t} = (NT)^{-1}\sum_{j=1}^{N}\sum_{t=1}^{T} \mathcal{SE}_{.jt}$.[8]

Strength of spatial dependence: One may want to assess how strongly a firm's innovation is dependent on other firms' R&D efforts. This is a form of spatial dependence, and its strength is denoted by the magnitude of ρ_2. A hypothesis test with the null $H_o:\ \rho_2 = 0$ can be used to assess the strength of spatial dependence. Rejecting H_o indicates that a firm's innovation is strongly dependent on other firms' R&D efforts.

Relevant forms of proximity between firms: By modelling the spatial matrix $\boldsymbol{w}$ using observable spatial covariates in $\boldsymbol{D}$ as a function of a parameter vector $\boldsymbol{\delta}$ of finite length k_δ, the proposed approach simultaneously allows several forms of connectivity between firms and determines their relevance statistically.[9] In a similar vein, the strength of time-variation, asymmetry, or both in $\boldsymbol{w}(\boldsymbol{\delta})$ is assessable by testing the significance (jointly or individually) of corresponding elements in $\boldsymbol{\delta}$.

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

[9] 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.

4 Extensions

4.1 Extensions of the model

The paper proposes an approach that requires observed spatial covariates, and the elements of the spatial matrix are non-negative.[10] Also, the number of elements in the spatial matrix to be modelled grows in the order of $O(N^2T)$. Consistent estimation of the spatial matrix requires a balanced panel, and this can result in the loss of some firms from the sample. I propose extensions to the baseline model eq. (3.1) to deal with these challenges though I do not explore the computational and inferential aspects.

Spatio-temporal dependence: Past outcomes of other firms may affect a firm's future outcomes. Such a phenomenon can be modelled by setting x_{it} to lagged outcome $y_{i,t-1}$.

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

$E(y_{it}|\mathbf{y}_{t-1}, \mathbf{z}_{it}, \boldsymbol{D}_t)$ 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 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, past outcome characterises a form of proximity and a function of it can be included 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 parameters.

Prior knowledge in spatial matrix: The researcher may have prior or expert knowledge that is desirable to incorporate in the estimation of the spatial matrix. An example is where the researcher has information on zero and non-zero elements of the spatial matrix but not

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

magnitudes. In the context of this paper, the researcher only has to model non-zero elements of the spatial matrix using the proposed approach.[11]

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, sect 5.2.1)). Group heterogeneity implicitly assumes that firms within the same group (e.g., industrial classification) are equally impacted by R&D of firms in other groups. Group heterogeneity is particularly helpful in dealing with unbalanced panel data as it avoids the loss of firms that results from constructing balanced out of unbalanced panels. The conditional expectation eq. (3.1), under group heterogeneity, is

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

where $w_{KLt}(\boldsymbol{\delta}) \equiv \dfrac{\exp(\boldsymbol{d}'_{KLt}\boldsymbol{\delta})}{\sum_{L \neq K} \exp(\boldsymbol{d}'_{KLt}\boldsymbol{\delta})}$, $\boldsymbol{d}_{KLt} \in \boldsymbol{D}_t$ contains spatial covariates specific to groups K and L, and $K \ni i$ means i belongs to group K. The baseline case obtains when there are N groups. Group heterogeneity leaves group membership across time unrestricted as long as group membership is observed. The researcher, however, needs to determine the criterion used to define group assignment, e.g., grouping firms by industry. Formulation 4.1 has a major advantage of not allowing the dimension of $\boldsymbol{w}_t(\boldsymbol{\delta})$ to grow as N grows *ad infinitum* if the number of groups is fixed.

4.2 Other possible empirical applications of the model

A notable aspect to my 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 the relevant spatial covariates as well as an assessment of time-variation and asymmetry in the spatial matrix $\boldsymbol{w}$. Empirically, my method

[11] See Lychagin, Pinkse, Slade, and Reenen (2016, Sect. IV) a discussion on this.

has applications beyond the study of R&D spillover effects on innovation. In the following, I provide other possible empirical applications of the proposed approach.

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, for example, geographic proximity (Syverson, 2011). Like Lychagin, Pinkse, Slade, and Reenen (2016), my framework is applicable to study this problem by jointly accounting for all plausible forms of proximity between firms (measured by the respective spatial covariates) in a production function framework. Unlike, e.g., Bloom, Schankerman, and Van Reenen (2013), the spatial matrix is determined within the model and the interaction between firms over time can vary and be asymmetric. The 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.

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 spatial econometrics literature. Markets are not perfectly segregated, and shocks to demand in a region impact demand behaviour in other regions. The proposed approach can contribute to this literature by allowing time-variation, asymmetry, and other spatial covariates (besides geographical contiguity) in determining the spatial matrix.

5 Estimation

The negative binomial 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 estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Estimation of eq. (3.4) seems 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., where one includes firm and year fixed effects with large N and large 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 available efficient routines for commonly used models like the negative binomial, linear, logit, or Poisson regressions.

In the following algorithm, I propose an iterative scheme that reduces direct minimisation 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 $\hat{\boldsymbol{\beta}}^{(l)} = \arg\min_{\boldsymbol{\beta} \in \boldsymbol{B}} \mathcal{Q}_n(\boldsymbol{\beta}, \hat{\boldsymbol{\delta}}^{(l-1)})$*, i.e., regress* $\mathbf{y}$ *on* $\boldsymbol{m}^{(l)}$

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

(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 number specified by the researcher, e.g., $\epsilon = 10^{-6}$. Constructing the design matrix involves plugging $\hat{\boldsymbol{\delta}}^{(l)}$ into $\boldsymbol{w}(\cdot)$ 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 to obtain $\hat{\boldsymbol{\beta}}^{(l)}$. In the case of the negative binomial model, the dispersion parameter η^2 is updated at step (c) as well. Minimising with respect to $\boldsymbol{\delta}$ at step (*d*) of each iteration saves much computational effort because the dimension of $\boldsymbol{\beta}$ generally far exceeds that of $\boldsymbol{\delta}$.[12]

[12] The dimension of $\boldsymbol{\beta}$ exceeds that of $\boldsymbol{\delta}$ by a factor of at least 8 in the empirical application (section 7).

6 Inference

6.1 Preliminary Concepts

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. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$.

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 $i = 1, \ldots, N$.*

Definition 2 (Conditional Stationarity). *The sequence of random variables $\{y_{it}\}_{t=1}^{T}$ for each $i \in \{1, \ldots, N\}$ 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 $\{y_{i\tau}\}_{\tau=t_1+r}^{t_k+r}$ conditioned on $\mathcal{F}$ almost surely for all $1 \leq t_k \leq (T-1)$, $1 \leq r \leq (T-t_k)$.*

The above definitions adapt definitions 4 and 5 in Rao (2009). The following assumptions are needed for the treatment of consistency and asymptotic normality of the estimator.

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 $\bar{\alpha}$, $\bar{\alpha}^{8.5} < \infty$, 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 $\lim\limits_{T\to\infty} \max\limits_{1\leq i\leq N} \sum\limits_{\iota=1}^{T} \alpha_i^{\mathcal{F}}(\iota) \leq \bar{\alpha}$. (c) For each $i \in \{1, \ldots, N\}$, $\{y_{it}\}_{t=1}^{T}$ conditional on $\mathcal{F}$ is stationary.*

Assumption 4(a) says outcomes y_{it} and $y_{i\tau}$, $i \neq j$ are conditionally independent; it does not impose serial independence on $\{y_{it}\}_{t=1}^{T}$ for each i. Also, it does not impose unconditional independence between the outcomes y_{it} and $y_{j\tau}$, $j \neq i$, but it requires conditional independence. The conditional strong-mixing assumption (assumption 4(b)) is needed to bound the sum of all conditional covariances of the outcome across $t = 1, \ldots, T$ for all pairs of firms $i, j \in \{1, \ldots, N\}$. It is not a strong assumption since (conditional) serial correlation is known to be weaker the farther apart outcomes (y_{it} and $y_{it+\iota}$), $\iota \geq 1$ are in time. In the context of the empirical application, assumption 4(c) says that the outcome (citation-weighted patent counts) of a firm i, conditional on observables including its own R&D, its characteristics, other firms' R&D, and year dummies, is stationary. Put together, assumption 4(a-c) is useful in establishing a bound on the sum of covariances of all pairs in $\{y_{it}\}_{i=1,t=1}^{N,T}$ conditional on $\mathcal{F}$.

The following proposition provides bounds on the sum of conditional covariances across all firms and time periods.

Proposition 1. *Under assumptions 4(a-c) and 2(b),* $\lim_{N,T\to\infty} \frac{1}{NT}\Big|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{t=1}^{T}\sum_{\tau\neq t} \operatorname{cov}(y_{it}, y_{j\tau}|\mathcal{F})\Big| \leq 8\bar{\alpha}\kappa_y$ *where the conditional covariance* $\operatorname{cov}(y_{it}, y_{j\tau}|\mathcal{F}) \equiv E(y_{it}y_{j\tau}|\mathcal{F}) - E(y_{it}|\mathcal{F})E(y_{j\tau}|\mathcal{F})$

6.2 Consistency

To proceed with the proof of consistency, let us make the following standard 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, it is shown 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 2(b) and 5,* $\mathcal{Q}_o(\boldsymbol{\beta}, \boldsymbol{\delta})$ *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 established in the following theorem.

Theorem 2 (Consistency)**.** *Under assumptions 2(a&b), 3(a-c), and 6,* $[\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}']' \xrightarrow{p} [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$.

6.3 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

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

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

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

Lemma 6. *Under assumptions 1, 4(a-c), 2(b), and 3(a-c), (a) the expectation of the score evaluated at $[\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ is zero, i.e., $\mathrm{E}[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] = \mathbf{0}$ and (b) the second moment of $(NT)^{1/2} s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$, $\mathrm{var}[(NT)^{1/2} s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$, is finite.*

Lemma 7. *Under the assumptions 2(b), 2(a-c), and 3(a-c), 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 $\mathrm{E}[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$. 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 satisfies $0 < ||\mathbf{a}^{it}||_2 < \infty$ 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(a) concerns the part of the hessian matrix corresponding to $\boldsymbol{\delta}$. It is introduced to ensure non-singularity in the hessian matrix. In the absence of assumption 6(b), the model reduces to a standard negative binomial model without spillover effects.

Lemma 8. *Under assumptions 2(b), 3(a-c), and 6, $\mathrm{E}[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$ is positive definite.*

In addition to the preceding lemmas, the following standard assumption is needed to establish the asymptotic normality of the estimator.

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 assumptions 2(a&b), 3(a-c), 7, and 6, $\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_\theta)$ where the variance-covariance matrix $\mathbf{V}_\theta \equiv \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$, $\mathbf{A}_o \equiv \mathrm{E}[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$, and $\mathbf{B}_o \equiv \mathrm{var}((NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o))$.*

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, I consider averages of the parameters of interest: average private effect, $\varrho_{n,\mathcal{PE}}(\boldsymbol{\beta}, \boldsymbol{\delta}) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})\boldsymbol{e}'_2\boldsymbol{\beta}$, 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 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 is given by $\varrho'_{n,\mathcal{SE}}(\boldsymbol{\beta},\boldsymbol{\delta}) = \varrho'_{n,\mathcal{PE}}(\boldsymbol{\beta},\boldsymbol{\delta}) + \varrho'_{n,\mathcal{S}p\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\delta})$.

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

Corollary 1. *Let $\varrho_n : \boldsymbol{\Theta} \to \mathbb{R}$ denote a parameter of interest, and $\varrho(\boldsymbol{\theta}) \equiv \lim_{N,T\to\infty} \varrho_n(\boldsymbol{\theta})$. Under assumptions 2(a&b), 3(a-c), 7, and 6, $\sqrt{NT}(\varrho_n(\hat{\boldsymbol{\theta}}) - \varrho(\boldsymbol{\theta}_o)) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \varrho'(\boldsymbol{\theta}_o)'\mathbf{V}_\theta \varrho'(\boldsymbol{\theta}_o))$ where $\varrho'(\cdot)$ denotes a $k_\theta \times 1$ the Jacobian matrix of $\varrho(\cdot)$.*

7 Empirical Application

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

7.1 Data

Data for the empirical analyses are an updated version of data used in Bloom, Schankerman, and Van Reenen (2013).[13] A balanced panel is constructed out of the original data set hence I do not replicate the results in Bloom, Schankerman, and Van Reenen (2013). The balanced panel comprises 217 firms in 21 two-digit Standard Industrial Classification

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

(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 outcomes (patent counts) but also the acknowledged relevance (citations). Pat_Cite has a mass point at zero; 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

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 measures using technology proximity $ln(SpTECH)$, product market proximity $ln(SpSIC)$, geographical proximity $ln(SpGEOG)$, and 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.

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), $ln(SpSIC)$ (in product market space), and $ln(SpGEOG)$ (using geographical location) are included for comparison. To control for firm size, log real sales revenue[14] $ln(Sales)$ are included. I include 3-digit SIC industry and year dummies in all specifications to control for unobserved heterogeneity.

7.2 Empirical model

I 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 handling outcome variables with mass points (see discussion in Berger, Stocker, and Zeileis (2017) and Huang and Tsyawo (2018)), the crux of this empirical application lies in modelling spatial interactions and estimating spillover effects in a possibly non-linear model. The model is applicable to the class of single-index models, and since this class

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

encompasses a number of methods used in the literature, for example, Poisson, the hurdle model, Tobit, Heckman selection, logit, distribution regression, and quantile regression, this paper's approach remains applicable.

The conditional mean is given by eq. (3.1), and 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, Bloom, Schankerman, and Van Reenen (2013)'s spillover measures $ln(SpTECH)$, $ln(SpSIC)$, and $ln(SpGEOG)$ are included for comparison. Standard errors are clustered at the firm level to allow for serial correlation.

Spatial covariates in $\boldsymbol{d}_{ijt} \in \boldsymbol{D}_t$ considered are the Jaffe measures of proximity 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 and j are in the same 3-digit industry classification ($dI(SIC3)$), a dummy equal to one if one of i, j (but not both) has zero lagged R&D stock $dI(R\&D)$, and a dummy equal to one if one of i, j (but not both) has lagged patent stock equal to zero $dI(Cite)$. The polynomial series terms on $R\&D_{i,t-2}$ and $R\&D_{j,t-2}$ are given by a vector comprising $\{R\&D_{i,t-2}^{l_i} R\&D_{i,t-2}^{l_j} : l_i = 0, \ldots, k-1, l_j = 1, \ldots, k - l_i\}$ where $k = 3$.

7.3 Results

The main results are summarised in tables 3 and 4. Since the main covariate of interest $ln(R\&D)_{t-1}$ 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 my 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 specifications (1) and (5) in table 4. Columns (2)-(8) in table 3 and columns (1)-(5) in table 4 are specifications using my approach with different combinations of spatial covariates in the weight function $w_{ijt}(\boldsymbol{\delta})$. Following the literature (e.g., Bloom, Schankerman, and Van Reenen (2013) and König, Liu, and Zenou (2019)), specification (8) in table 3 treats R&D as endogenous and uses R&D tax (both State and Federal) as instruments.[15] 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 of elements in $\boldsymbol{\delta}$ corresponding to spatial covariates (subscript of 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 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 my approach

[15] See Bloom, Schankerman, and Van Reenen (2013, sect. 5.1) for details.

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)	Neg. Bin (8)
Panel A: ($\boldsymbol{\beta}$)								
ρ_1	0.155	0.187	0.176	0.186	0.181	0.170	0.157	0.194
	(0.044)	(0.042)	(0.043)	(0.042)	(0.041)	(0.041)	(0.045)	(0.048)
ρ_2		0.048	-1.331	1.566	0.213	0.319	0.325	0.088
		(0.402)	(1.173)	(0.534)	(0.070)	(0.067)	(0.067)	(0.042)
$ln(SpTECH)_{t-1}$	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	0.559
	(0.034)	(0.035)	(0.033)	(0.035)	(0.034)	(0.033)	(0.033)	(0.032)
Panel B: ($\boldsymbol{\delta}$)								
χ^2_{dTECH}		0.210				0.94	1.000	0.017
df, $P(>\chi^2)$		1,(0.650)				1,(0.330)	1,(0.310)	1, (0.900)
χ^2_{dSIC}			122.200			10.600	13.400	0.077
df, $P(>\chi^2)$			1,(0.000)			1,(0.001)	1,(0.000)	1,(0.780)
χ^2_{dGEOG}				34.100		14.000	12.000	0.940
df, $P(>\chi^2)$				1,(0.000)		1,(0.000)	1,(0.001)	1, (0.330)
$\chi^2_{dR\&D_{t-2}}$					190.700	50.000	49.100	24.500
df, $P(>\chi^2)$					6,(0.000)	6,(0.000)	6,(0.000)	6, (0.000)
$\chi^2_{dI(SIC3)}$						1.000	0.910	1.100
df, $P(>\chi^2)$						1,(0.310)	1, (0.340)	1, (0.300)
$\chi^2_{dI(R\&D)}$						0.76	0.620	0.310
df, $P(>\chi^2)$						1,(0.380)	1,(0.430)	1, (0.580)
$\chi^2_{dI(Cite)}$						5.000	5.200	0.270
df, $P(>\chi^2)$						1,(0.025)	1,(0.022)	1, (0.600)

Notes: All specifications above contain firm and year fixed effects. The outcome variable is citation-weighted patent counts (Pat_Cite_t). Number of firm-year observations: 5859. Panel A reports coefficients and standard errors (in parentheses). Standard errors allow for serial correlation via clustering at the firm level. The variable **x** is lagged log R&D $ln(R\&D)_{t-1}$. In panel B, the spatial covariates (with respective Wald-statistics) considered are: Jaffe measures in technology ($dTECH$), product market ($dSIC$), and geographical proximity ($dGEOG$). Others are 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 $i,\ j$) 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 $dI(Cite)$. $P(>\chi^2)$ (in parentheses) denotes the p-value and df equals the length of the corresponding parameter vector. A two-step control function is implemented in specification (8) with first three terms of the polynomial expansion of the residual from the first stage. The first stage F-statistic of the lagged log R&D stock on R&D tax prices (as instrumental variables) is 671.8.

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 (8) in table 3 and specifications (2)-(5) in table 4 gives the spillover elasticity of $R\&D$ on firm innovation using the proposed approach. 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. $dTECH$, $dGEOG$, and $dR\&D_{t-2}$ in specifications (2), (3)-(5) are spatial covariates associated with positive ρ_2. Save specifications (2),(3), and (8) 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 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, and this 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 spatial covariates in $\boldsymbol{d}_{ijt}$ is indicative of support against a null hypothesis that each element of $\boldsymbol{w}(\boldsymbol{\delta})$ is equal to $1/(N-1)$ where N is the number of firms.[16] Rejecting this hypothesis does, however, not indicate the relevance of the spatial covariate vis-à-vis the outcome is concerned. It ought to be coupled with a significant 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 specification (3), $dI(SIC3)$ in table 4 specification (2), and $dI(R\&D)$ table 4 specification (3) 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

[16] $w_{ijt}(\boldsymbol{\delta}) \equiv \dfrac{\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\sum_{j\neq i}^{N}\exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})} = 1/(N-1)$ 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)
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 contain firm and year fixed effects. The outcome variable is citation-weighted patent counts (Pat_Cite_t). Number of firm-year observations: 5859. Panel A reports coefficients and standard errors (in parentheses). Standard errors allow for serial correlation via clustering at the firm level. The variable $\mathbf{x}$ is lagged log R&D $ln(R\&D)_{t-1}$. In panel B, the spatial 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 $dI(Cite)$. $P(>\chi^2)$ (in parentheses) denotes the p-value and df equals the length of the corresponding parameter vector.

not significant at any conventional level. Elements of $\boldsymbol{\delta}$ 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 positive R&D spillover effects on firm innovation and the relevance of $dSIC$, $dGEOG$, $dR\&D_{t-2}$ and $dI(Cite)$ as spatial covariates. Specification (8) implements a control function approach to control for endogeneity in R&D stock, as in for example, Bloom, Schankerman, and Van Reenen (2013). The spillover effect is positive and significant at the 5% level. It is, however, smaller in magnitude vis-à-vis specifications (6) and (7). In the set of relevant spatial covariates, $dR\&D_{t-2}$ and $dI(Cite)$ are time-varying and $dR\&D_{t-2}$ is asymmetric. This result indicates that the interaction structure between firms with respect to innovation evolves through time. Asymmetry in the spatial matrix implies some firms generate more spillover effects than others and that tax credit/subsidy policies that target firms equally are not optimal from a social perspective.

In sum, estimation using specification (6) in table 3 (the preferred specification) 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 include 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 often assumed measure of proximity (e.g., geographic contiguity), and they induce asymmetry and time variation in the spatial matrix.

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, I address the problem of (partly) unknown spatial matrices in a study that quantifies R&D spillover effects on firm innovation. The proposed approach is robust to misspecification of the spatial matrix, and it simultaneously controls for several forms of proximity between firms by using a vector of spatial covariates. 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 applies to the class of single-index models. The approach tackles the problem of metric ambiguity by accommodating several spatial covariates simultaneously and assesses their relevance via hypothesis tests. I provide identification results and establish asymptotic properties of the MLE estimator.

The empirical application considered studies R&D spillovers on innovation in a knowledge production framework. The importance of the approach is confirmed by the sensitivity of spillover effect estimates to the choice of spatial covariate and spatial matrix by extension. The 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, a main finding is that product market proximity, geographic proximity, past R&D, and past patent citations are relevant forms of connectivity between firms with respect to innovation. The result on relevant spatial covariates confirms 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. This result indicates that while spillovers are important from a policy-perspective, it is worthy of note that their channels of transmission, besides geographic and product market proximity, may be multidimensional and may not always tied to often perceived notions of closeness.

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* (2019), 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] König, Michael D, Xiaodong Liu, and Yves Zenou. “R&D Networks: Theory, Empirics, and Policy Implications”. *Review of Economics and Statistics* 101.3 (2019), pp. 476–491.

[23] 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.

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

[25] 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.

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

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

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

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

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

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

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

[33] 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 lemmas

In the following lemma, I 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 2, 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)* $\sup_{\substack{\boldsymbol{\beta}\in\boldsymbol{B}\\ \boldsymbol{\delta}\in\boldsymbol{\Delta}}} \left|\left|\mu_{it}(\boldsymbol{\beta},\boldsymbol{\delta})\right|\right|_2 \leq 2\kappa_\beta\kappa_d\kappa_w\kappa_z$ *where* $\mu_{it}(\boldsymbol{\beta},\boldsymbol{\delta}) \equiv \frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}'}$*, 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 2(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 2(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 2(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, $||\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}||_2 = ||\rho_2\sum_{j\neq i} x_{jt}\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}||_2 \leq |\rho_2|\sum_{j\neq i}|x_{jt}| \cdot$

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

Lemma 10. *Under assumption 3(b), $\boldsymbol{w}(\boldsymbol{\delta}) \neq \boldsymbol{w}(\boldsymbol{\delta}_o)$ for any $\boldsymbol{\delta} \in \boldsymbol{\Delta}$, $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$.*

Proof of lemma 10. Applying the mean-value theorem, $w_{ijt}(\boldsymbol{\delta}) - w_{ijt}(\boldsymbol{\delta}_o) = \frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}'}|_{\boldsymbol{\delta}=\bar{\boldsymbol{\delta}}}(\boldsymbol{\delta} - \boldsymbol{\delta}_o)$ where $\bar{\boldsymbol{\delta}}$ is a vector between $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_o$. Using the expression of $\frac{\partial w_{ijt}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}'}$ from section 6.3, $w_{ijt}(\boldsymbol{\delta}) - w_{ijt}(\boldsymbol{\delta}_o) = w_{ijt}(\bar{\boldsymbol{\delta}})\Psi_{ijt}(\bar{\boldsymbol{\delta}})(\boldsymbol{\delta} - \boldsymbol{\delta}_o)$ where $\Psi_{ijt}(\boldsymbol{\delta}) \equiv [\boldsymbol{d}_{ijt} - \sum_{l \neq i} w_{ilt}(\boldsymbol{\delta})\boldsymbol{d}_{ilt}]'$. For $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$, $\Psi_{ijt}(\bar{\boldsymbol{\delta}})(\boldsymbol{\delta} - \boldsymbol{\delta}_o) \neq 0$ a.s., and in matrix notation, $\boldsymbol{w}(\boldsymbol{\delta}) - \boldsymbol{w}(\boldsymbol{\delta}_o) = \boldsymbol{w}(\bar{\boldsymbol{\delta}}) \circ \boldsymbol{\Psi} \neq \boldsymbol{0}$ where $\circ$ denotes the Hadamard product, $\Psi_{ijt}(\bar{\boldsymbol{\delta}})(\boldsymbol{\delta} - \boldsymbol{\delta}_o)$ is the (i,j)'th element of $\boldsymbol{\Psi}_t$, $t = 1, \ldots, T$, and $\boldsymbol{\Psi}$ is the block-diagonal matrix $\text{diag}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_T)$. □

B Proofs of lemmas and theorems

Proof of lemma 1.
First, $Nw_{ijt}(\boldsymbol{\delta}) = \dfrac{N \exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}{\sum_{j \neq i}^{N} \exp(\boldsymbol{d}'_{ijt}\boldsymbol{\delta})}$ is bounded a.s. for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ because

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

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 2(a).

Second, from the first part, $Nw_{ijt}(\boldsymbol{\delta}) \leq \kappa_w$. This implies that for any $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ and $j \in \{1, \ldots, N\}$, $j \neq i$, and $t \in \{1, \ldots, T\}$,

$$\sum_{i=1}^{N} w_{ijt}(\boldsymbol{\delta}) \leq \max_{\substack{1 \leq j \leq N \\ 1 \leq t \leq T}} \sum_{i=1}^{N} w_{ijt}(\boldsymbol{\delta}) \leq N \sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \max_{\substack{1 \leq j \leq N \\ 1 \leq t \leq T}} w_{ijt}(\boldsymbol{\delta}) \leq N\frac{\kappa_w}{N} = \kappa_w$$

By definition, the maximum column sum is $\max_{\substack{1 \leq j \leq N \\ 1 \leq t \leq T}} \sum_{i=1}^{N} w_{ijt}(\boldsymbol{\delta})$ because $\boldsymbol{w}(\boldsymbol{\delta})$ is block-diagonal.

$\square$

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 3(a), $\boldsymbol{z}'\boldsymbol{z}$ is invertible. By assumption 3(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, one only needs to show that the Shur complement of $\boldsymbol{z}'\boldsymbol{z}$ in $\boldsymbol{m}'\boldsymbol{m}$, i.e., $S = \mathbf{x}'\boldsymbol{w}'\boldsymbol{w}\mathbf{x} - \mathbf{x}'\boldsymbol{w}'\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{w}\mathbf{x}$ is positive.

Factor S as $S = \mathbf{x}'\boldsymbol{w}'[\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}']\boldsymbol{w}\mathbf{x}$. $\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'$ 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 $\mathbf{x} \neq \mathbf{0}$ (assumption 3(a)), it implies $S = \mathbf{x}'\boldsymbol{w}'[\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'][\mathbf{I}_{NT} - \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}']\boldsymbol{w}\mathbf{x} = \sum_{i=1}^{N}\sum_{t=1}^{T}\zeta_{it}^2 > 0$ 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.

From assumption 2(b), $\sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} ||m_{it}(\boldsymbol{\delta})||_2 \leq \kappa_m$. It follows that $\frac{1}{NT}||\boldsymbol{m}'\boldsymbol{m}||_2 \leq \frac{1}{NT}||\boldsymbol{m}||_2^2 \leq \kappa_m^2 < \infty$ for all $\boldsymbol{\delta} \in \boldsymbol{\Delta}$. $\square$

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, i.e., $\log u \leq u - 1$. Under assumption 2(b), $\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}) \leq \exp(|m_{it}(\boldsymbol{\delta})\boldsymbol{\beta}|) \leq \sqrt{\kappa_y} < \infty$.

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 (assumption 1).

$E(q_2) \equiv |(\eta^{-2}+y_{it})(\eta^{-2}-1)| = |\eta^{-2}-1|E(\eta^{-2}+y_{it}) = |\eta^{-2}-1|(\eta^{-2}+\mathrm{E}[E(y_{it}|\cdot)]) = |\eta^{-2}-1|(\eta^{-2}+\mathrm{E}[\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)]) \leq |\eta^{-2}-1|(\eta^{-2}+\sqrt{\kappa_y}) < \infty$. The first equality follows because $y_{it} \geq 0$, and $|(\eta^{-2}+y_{it})| = (\eta^{-2}+y_{it})$. $E(y_{it}|\cdot) \equiv \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$ by definition. The upper bound obtains using assumption 2(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 2(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 2(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 3(b)) which implies that $\boldsymbol{w}(\boldsymbol{\delta}) \neq$

$\boldsymbol{w}(\boldsymbol{\delta}_o)$ if $\boldsymbol{\delta} \neq \boldsymbol{\delta}_o$ a.s (lemma 10). Since $\mathbf{x} \neq \mathbf{0}$ (assumption 3(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 $[\boldsymbol{\beta}', \boldsymbol{\delta}']' \neq [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$.

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 $\boldsymbol{\beta} \neq \boldsymbol{\beta}_o$ a.s., and $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ is identified.

□

Proof of theorem 1. Under assumptions 2(b) and 3(a-c), $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o', \boldsymbol{\delta}_o']'$ is identified (lemma 2). Also, under assumption 2(b), $E|\mathcal{Q}_n(\boldsymbol{\beta}, \boldsymbol{\delta})| < \infty$ (lemma 3). 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 2(b). The conclusion follows from 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, $E|q_{it}(\boldsymbol{\beta},\boldsymbol{\delta})| \leq |\eta^{-2}\log\eta^{-2}| + |\eta^{-2}-1|(\eta^{-2}+\sqrt{\kappa_y}) + \eta^{-2}\sqrt{\kappa_y} + \kappa_y + 1/2\sqrt{\kappa_y}\log\kappa_y < \infty$ 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 assumptions 2(a&b), 3(a-c), and 6, $\mathcal{Q}_o(\boldsymbol{\beta},\boldsymbol{\delta})$ is uniquely minimised at $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o',\boldsymbol{\delta}_o']'$ (theorem 1).

Second, under assumptions 2(b) and 5, $\mathcal{Q}_o(\boldsymbol{\beta},\boldsymbol{\delta})$ is continuous, and $\sup_{\boldsymbol{\beta}\in\boldsymbol{B},\boldsymbol{\delta}\in\boldsymbol{\Delta}} |\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta}) - \mathcal{Q}_o(\boldsymbol{\beta},\boldsymbol{\delta})| \xrightarrow{p} 0$ (lemma 5).

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

Proof of lemma 6. Part (a):

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

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

Part (b):

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

The first equality follows because $\mathrm{E}[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] = \mathbf{0}$ (lemma 6(a)). The last line applies the law of iterated expectations and uses assumption 4(a), noting that $\mathrm{E}[s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o) s_{j\tau}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)'|\cdot] = \mathbf{0}_{k_\theta \times k_\theta}$ for $i \neq j$ (see also proposition 1), i.e., considering the part dependent on $\{y_{it}\}_{i=1,t=1}^{N,T}$, $\mathrm{E}[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - y_{it})(\exp(m_{j\tau}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - y_{j\tau})|\cdot] = E(y_{it}y_{j\tau}|\cdot) - \exp((m_{it}(\boldsymbol{\delta}_o) + m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) = E(y_{it}|\cdot)E(y_{j\tau}|\cdot) - \exp((m_{it}(\boldsymbol{\delta}_o) + m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) = \exp((m_{it}(\boldsymbol{\delta}_o) + m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) - \exp((m_{it}(\boldsymbol{\delta}_o) + m_{j\tau}(\boldsymbol{\delta}_o))\boldsymbol{\beta}_o) = 0$ for $i \neq j$.

First, consider

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

where $\boldsymbol{G}_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o) \equiv \begin{bmatrix} m_{it}(\boldsymbol{\delta}_o)' m_{it}(\boldsymbol{\delta}_o) & m_{it}(\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o) \\ \mu_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)' m_{it}(\boldsymbol{\delta}) & \mu_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o) \end{bmatrix}$ 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 $\mathrm{E}[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - y_{it})^2|\cdot] = E(y_{it}^2|\cdot) - \exp(2m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$. 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), $\mathrm{E}[(\exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) - y_{it})^2|\cdot] = \exp(m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o) + \eta^2\exp(2m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o)$. Simplifying terms, $\mathrm{E}[s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o) s_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)'|\cdot] = \frac{1}{(\eta^{-2} + \exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it}(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$.

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

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

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

The next step is to establish bounds on terms. Note that $||\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2 = ||m_{it}(\boldsymbol{\delta}_o)' m_{i\tau}(\boldsymbol{\delta}_o)||_2^2 + 2||m_{it}(\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2 \leq ||m_{it}(\boldsymbol{\delta}_o)||_2^4 + 2||m_{it}(\boldsymbol{\delta}_o)||_2^2 \cdot ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^4 = (||m_{it}(\boldsymbol{\delta}_o)||_2^2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2)^2$ from which it follows that $||\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 \leq ||m_{it}(\boldsymbol{\delta}_o)||_2^2 + ||\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2^2 \leq \kappa_m^2 + 4\kappa_\beta^2\kappa_d^2\kappa_w^2\kappa_z^2$. An analogous argument holds for $||\boldsymbol{G}_{it\tau}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)||_2 \leq \kappa_m^2 + 4\kappa_\beta^2\kappa_d^2\kappa_w^2\kappa_z^2$. For the first part, the following bound holds.

$$\Big|\Big|E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}[s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)s_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)'|\cdot]\Big]\Big|\Big|_2 = \Big|\Big|E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{(\eta^{-2} + \exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}\boldsymbol{G}_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)\Big]\Big|\Big|_2$$
$$\leq \eta^2(\kappa_m^2 + 4\kappa_\beta^2\kappa_d^2\kappa_w^2\kappa_z^2)$$

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,

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

The second equality follows from assumption 4(c), and the last inequality uses proposition 1. Both parts are bounded. Applying the triangle inequality completes the proof. □

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

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

$$
\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(\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)\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})'\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})(\kappa_\beta+2)
\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 using assumption 2(a-c).

$$
\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]\\
&\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}\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]\\
&\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}\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}\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|\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(\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)\Big|\Big|\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\delta}\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|\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]\\
&\leq 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]\\
&\leq \kappa_\beta\kappa_d\kappa_w\kappa_z(\eta^{-2}+\sqrt{\kappa_y})(2\kappa_\beta\kappa_d\kappa_w\kappa_z+5\kappa_d)
\end{aligned}
$$

Let $\bar{\kappa}\equiv\max\{\kappa_\beta,\kappa_d,\kappa_m,\kappa_w,\kappa_y,\kappa_z\}$, and let the notation $\xi(\bar{\kappa})\propto\bar{\kappa}^r$, $r\geq 0$ denote a function ξ of up to the r'th polynomial term of $\bar{\kappa}$. Then, $E\Big[\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big|\Big|_2^2\Big]\leq(\frac{\kappa_m^2}{2}(\eta^{-2}+\sqrt{\kappa_y}))^2\propto\bar{\kappa}^5$, $E\Big[\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\Big|\Big|_2^2\Big]\leq(\kappa_d\kappa_w\kappa_z(\eta^{-2}+\sqrt{\kappa_y})(\kappa_\beta+2))^2\propto\bar{\kappa}^9$, and $E\Big[\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2^2\Big]\leq(\kappa_\beta\kappa_d\kappa_w\kappa_z(\eta^{-2}+\sqrt{\kappa_y})(2\kappa_\beta\kappa_d\kappa_w\kappa_z+5\kappa_d))^2\propto\bar{\kappa}^{17}$. Note that $E||H_n(\boldsymbol{\beta},\boldsymbol{\delta})||_2^2=E\Big[\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big|\Big|_2^2+2\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\Big|\Big|_2^2+\Big|\Big|\frac{\partial\mathcal{Q}_n(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\Big|\Big|_2^2\Big]\leq(\frac{\kappa_m^2}{2}(\eta^{-2}+\sqrt{\kappa_y}))^2+2(\kappa_d\kappa_w\kappa_z(\eta^{-2}+\sqrt{\kappa_y})(\kappa_\beta+2))^2+(\kappa_\beta\kappa_d\kappa_w\kappa_z(\eta^{-2}+$

$\sqrt{\kappa_y})(2\kappa_\beta\kappa_d\kappa_w\kappa_z + 5\kappa_d))^2 \propto \bar{\kappa}^{17}$. By Jensen's inequality, $(E||H_n(\boldsymbol{\beta},\boldsymbol{\delta})||_2)^2 \leq E||H_n(\boldsymbol{\beta},\boldsymbol{\delta})||_2^2$ which implies $E||H_n(\boldsymbol{\beta},\boldsymbol{\delta})||_2 \propto \bar{\kappa}^{8.5} < \infty$.

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 under assumption 2(a-c). 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. □

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

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

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 $\dfrac{\eta^{-2}}{(1+\eta^{-2}\exp(-m_{it}(\boldsymbol{\delta}_o)\boldsymbol{\beta}_o))}$.

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

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 $\mu_{it}(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$. The third equality follows from applying the LIE, noting that $E\Big[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta^{-2}(\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})-E(y_{it}|\cdot))}{\eta^{-2}+\exp(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}\frac{\partial(m_{it}(\boldsymbol{\delta})\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}'}\Big]\Big|_{\substack{\boldsymbol{\beta}=\boldsymbol{\beta}_o\\ \boldsymbol{\delta}=\boldsymbol{\delta}_o}} = \mathbf{0}$

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

The second part of the term after the second equality is zero using an argument analogous to that of part (b) above.

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

To prove positive definiteness of $\mathrm{E}[H_n(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)]$, first note that $\mathrm{E}[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a]$ is invertible using that $\mathrm{E}[\frac{1}{NT}\boldsymbol{m}'\boldsymbol{m}]$ 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 $\mathrm{E}[H_n(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)]$ given by

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

where $\hat{\boldsymbol{\mu}}_a = \frac{1}{NT}\boldsymbol{\mu}_a'(\mathbf{I}_{NT} - \boldsymbol{m}_a(\mathrm{E}[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}\boldsymbol{m}_a')$ 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 $\rho_{o,2}$ denotes the third element in $\boldsymbol{\beta}_o$. It is necessary to show that $(\sum_{j\neq i}(\rho_{o,2}w_{ijt}(\boldsymbol{\delta}_o)(\sum_{\substack{l\neq i\\ l\neq j}}x_{lt}))^2)^{1/2} <$

∞ a.s. $|\rho_{o,2}| \leq \kappa_\beta < \infty$ by assumption 5 where $\kappa_\beta \equiv \sup_{\boldsymbol{\beta}\in\boldsymbol{B}} ||\boldsymbol{\beta}||_\infty$. $w_{ijt}(\boldsymbol{\delta}_o) \leq \frac{\kappa_w}{N}$ for all $i, j = 1, \ldots, N$, $j \neq i$, by lemma 1. $|\sum_{\substack{l\neq i\\ l\neq j}} x_{lt}| \leq \sum_{\substack{l\neq i\\ l\neq j}} |x_{lt}| \leq (N-2)\kappa_z$ 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 $\mathbf{x} \neq \mathbf{0}$ assumption 3(a). Since $\boldsymbol{D}_a(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ 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., rank$(\boldsymbol{\mu}_a) = k_\delta$. Note that $\boldsymbol{m}_a(\mathrm{E}[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}\boldsymbol{m}_a'$ 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 $\mathbf{I}_{NT} - \boldsymbol{m}_a(\mathrm{E}[\frac{1}{NT}\boldsymbol{m}_a'\boldsymbol{m}_a])^{-1}\boldsymbol{m}_a'$ is $NT - k_\beta$. $NT - k_\beta - k_\delta > 0$ for any estimable model hence it follows that the rank$(\hat{\boldsymbol{\mu}}_a) = k_\delta$ and the Schur complement $\mathrm{E}[\hat{\boldsymbol{\mu}}_a'\hat{\boldsymbol{\mu}}_a]$ 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 2(a&b), 3(a-c), and 6,, $\hat{\boldsymbol{\theta}} \equiv [\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}']'$ is consistent (theorem 2). By lemma 6(a), $\mathrm{E}[s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)] = \mathbf{0}$ and under assumptions 1, 4(a-c), 2(b), and 3(a-c), the second moment of $(NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)$ is finite (lemma 6(b)). Under the assumptions 2(b), 2(a-c), and 3(a-c), the hessian matrix $H_n(\boldsymbol{\beta}, \boldsymbol{\delta})$ is bounded in absolute value for all $[\boldsymbol{\beta}', \boldsymbol{\delta}']' \in \boldsymbol{\Theta}$ (lemma 7). Under assumptions 2(b), 3(a-c), and 6, $\mathrm{E}[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$ is positive definite (lemma 8). 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 $\mathbf{V}_\theta \equiv \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$, $\mathbf{A}_o \equiv \mathrm{E}[H_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o)]$, and $\mathbf{B}_o \equiv \mathrm{var}((NT)^{1/2}s_n(\boldsymbol{\beta}_o, \boldsymbol{\delta}_o))$. □

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 vector between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$. $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$ (theorem 2) and $\bar{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_o$ because it is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$. $\varrho'(\bar{\boldsymbol{\theta}}) \xrightarrow{p} \varrho'(\boldsymbol{\theta}_o)$ by the continuous mapping theorem. It follows from theorem 3 that $\sqrt{NT}(\varrho(\hat{\boldsymbol{\theta}})-\varrho(\boldsymbol{\theta}_o)) = \varrho'(\boldsymbol{\theta}_o)\sqrt{NT}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)+o_p(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \varrho'(\boldsymbol{\theta}_o)'\mathbf{V}_\theta\varrho'(\boldsymbol{\theta}_o))$. $\square$