Estimating spillovers from sampled links

Kieran Marray

VU Amsterdam, Tinbergen Institute

k.j.marray@vu.nl

Econometric Society European Meeting, Summer 2024

Empirical papers on spillovers often observe a **sampled network** containing a **subset (undersampling)** or **superset (oversampling)** of true links between individuals.

Empirical papers on spillovers often observe a **sampled network** containing a **subset (undersampling)** or **superset (oversampling)** of true links between individuals.

Fixed choice surveys – ask participants to name M friends/name friends from list of M others (e.g Harris 2009, Conley & Udry 2010, Oster & Thornton 2012, Banerjee et al. 2013) – undersample links if maximum degree is greater than M

Proxying links by proximity in some space – assume all individuals in same classroom/technology class/location/ethnicity interact (e.g Manski 1993, Miguel & Kremer 2004, Beaman 2011, Bloom et al. 2013, Carrell et al. 2013).

Estimate spillover effects by regressing **spillovers on sampled network** on outcomes.

Oster & Thornton (2012) "In addition, given the randomization, we are able to obtain an unbiased estimate of the impact of additional treatment friends even if we do not observe all of an individual's friends."

Here, I:

1. show that spillover estimates from undersampled/oversampled links are (economically significantly) biased, often upwards – sampling induces dependence between observed and unobserved spillovers ▸▸ size of bias.

Here, I:

1. show that spillover estimates from undersampled/oversampled links are (economically significantly) biased, often upwards – sampling induces dependence between observed and unobserved spillovers size of bias.
2. show how researchers can construct unbiased estimators or assess robustness of estimates to number of missing links without conditioning on network formation model, and

Here, I:

1. show that spillover estimates from undersampled/oversampled links are (economically significantly) biased, often upwards – sampling induces dependence between observed and unobserved spillovers size of bias.
2. show how researchers can construct unbiased estimators or assess robustness of estimates to number of missing links without conditioning on network formation model, and
3. apply to re-estimates of spillovers of climate shocks on firm-level production networks from Barrot & Sauvagnat (2016) – estimates are 1.5-2 times too large due to sampling bias.

Empirical literature
Education Rapoport & Horvath (1961), Harris (2009), Calvó-Armengol et al. (2009), Carrell et al. (2013). **Development** Miguel & Kremer (2004), Banerjee et al. (2013), Oster & Thornton (2012). **Innovation** Jaffe (1986), Foster & Rosenzweig (1995), Bloom et al. (2013). **Labour** Munshi (2003), Beaman (2011), ...

Econometric literature
Construct unbiased estimates without throwing away data (Chandrasekhar & Lewis 2016) or conditioning on a specific network formation model (Breza et al. 2020, Herstad 2023, Yauck 2022, Zhang 2023, Hseih et al. 2024, e.g). Nest cases in (Griffith 2022, Lewbel et al. 2022) for cases presented there.

Closely related to problem of endogenous exposure to exogenous shocks in design-based estimation of causal effects (Borusyak & Hull 2023) – can construct unbiased estimates without knowing counterfactual exposure process.

Sampling

True network G^*, sampled network G, unobserved adjacency matrix B.

Treatment variable X, outcome $Y_i = h((G^*X)_i)$ (Aronow & Samii 2021).

$$G^* = G + B \implies G^*X = GX + BX$$
$$\implies (GX)_i = \begin{cases} (G^*X)_i - (BX)_i & \text{if incorrectly sampled,} \\ (G^*X)_i & \text{else.} \end{cases}$$

Undersampling $\implies \sum_j B_{ij}$ bigger when $\sum_j G_{ij}$ bigger; oversampling $\implies \sum_j B_{ij}$ bigger when $\sum_j G_{ij}$ smaller.

Therefore even if X is i.i.d. often

$$E(BX) \neq 0, \text{plim } N^{-1}(GX)'BX \neq 0$$

sampling induces dependence between observed and unobserved spillovers – plus $Y_i = h((G^*X)_i)$ gives non-classical measurement error.

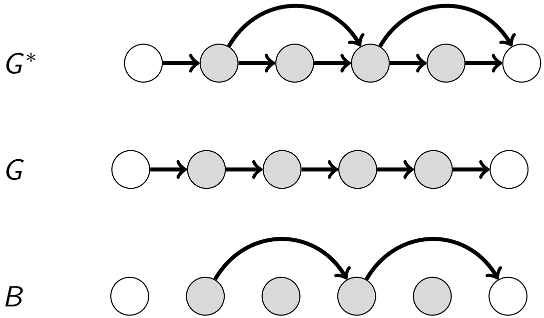


Figure: Undersampling on a line network

$$GX = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, BX = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

$$N^{-1}(GX)'BX = \frac{1}{6}\begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

$$= \frac{1}{3} \neq 0.$$

Formal example

Linear models

Sampling biases linear regression estimates. Assume: A1) Lindenberg conditions, A2) $BX \perp \epsilon | G^* X$.

DGP: $Y = \alpha + X\gamma + G^* X\beta + \epsilon$. Sample analogue: $Y = \alpha + X\gamma + GX\beta$.

Estimator: $\hat{\beta}^{\text{OLS}} = ((GX)' GX)^{-1}(GX)' Y$.

$$E(\hat{\beta}^{\text{OLS}}) = \beta + \beta E(((GX)'(GX))^{-1}((GX)' BX)).$$

Sampling biases linear regression estimates. Assume: A1) Lindenberg conditions, A2) $BX \perp \epsilon | G^* X$.

DGP: $Y = \alpha + X\gamma + G^* X\beta + \epsilon$. Sample analogue: $Y = \alpha + X\gamma + GX\beta$.

Estimator: $\hat{\beta}^{\text{OLS}} = ((GX)' GX)^{-1}(GX)' Y$.

$$E(\hat{\beta}^{\text{OLS}}) = \beta + \beta E(((GX)'(GX))^{-1}((GX)'BX)).$$

Proposition

Assume A1), A2).

$$\text{Bias: } E(\hat{\beta}^{\text{OLS}} - \beta) = E(A^{-1}(GX)'BX\beta)$$
$$\text{Inconsistency: } \operatorname{plim} \hat{\beta}^{\text{OLS}} - \beta = \operatorname{plim} A^{-1}((GX)'BX)\beta,$$

where $A := (GX)'(GX)$.

We can rescale OLS estimates by **the dependence between observed and unobserved spillovers**.

Proposition

Assume A1), A2). The estimator

$$\hat{\beta} = (I + \eta)^{-1} \hat{\beta}^{\text{ OLS}} \text{ where } \eta = E(A^{-1}(GX)'BX) \tag{1}$$

is an unbiased and consistent estimator of β

$$E(\hat{\beta}) = \beta, \text{ and plim } \hat{\beta} = \beta.$$

We can rescale OLS estimates by **the dependence between observed and unobserved spillovers**.

Proposition

Assume A1), A2). The estimator

$$\hat{\beta} = (I + \eta)^{-1} \hat{\beta}^{\text{ OLS}} \text{ where } \eta = E(A^{-1}(GX)' BX) \tag{1}$$

is an unbiased and consistent estimator of β

$$E(\hat{\beta}) = \beta, \text{ and plim } \hat{\beta} = \beta.$$

In practice, make **conditional independence assumption** (that we will relax later)
A3) $(G^*, B) \perp X$ (e.g RCT on network) – then $E((GX)'BX) = Nd^G d^B \bar{X}^2$.

Only need **one more survey question – 'how many friends do you have?'**

What if you cannot get information on the true mean degree?

What if you cannot get information on the true mean degree?

$$\textbf{Robustness to missingness}: \hat{\beta}^{\text{ OLS}} > \tau \iff d^B < \left(\frac{1}{NA^{-1}\bar{X}^2 d^G}\right)\frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}.$$

What if you cannot get information on the true mean degree?

Robustness to missingness: $\hat{\beta}^{\text{ OLS}} > \tau \iff d^B < \left(\frac{1}{NA^{-1}\bar{X}^2 d^G}\right)\frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}.$

Bounds: $d^B \in [d^B_{\text{min}}, d^B_{\text{max}}] \implies \beta \in \left[\frac{\hat{\beta}^{\text{ ols}}}{I + \eta(d^B_{\text{max}})}, \frac{\hat{\beta}^{\text{ ols}}}{I + \eta(d^B_{\text{min}})}\right]$

Extensions

Might instead fit the non-linear model (Blume et al. 2015)

$$\text{DGP: } Y = \lambda G^* Y + X\beta + \epsilon, \text{ sample analogue } Y = \lambda GY + X\beta \tag{2}$$

by two-stage least-squares using treatment of sampled friends of sampled friends as instruments (Kelejian & Prucha 1998).

Might instead fit the non-linear model (Blume et al. 2015)

$$\text{DGP: } Y = \lambda G^* Y + X\beta + \epsilon, \text{ sample analogue } Y = \lambda GY + X\beta \tag{2}$$

by two-stage least-squares using treatment of sampled friends of sampled friends as instruments (Kelejian & Prucha 1998).

Proposition

Make A2) and standard SAR assumptions. Let P denote a projection matrix, $Z^{\text{2SLS}} = [GY, X]$, $H^{\text{2SLS}} = [X, GX, G'GX, ...]$. The two-stage least-squares estimator

$$\hat{\theta}^{\text{ 2SLS}} = \begin{pmatrix} \hat{\lambda}^{\text{ 2SLS}} \\ \hat{\beta}^{\text{ 2SLS}} \end{pmatrix} = (Z' P_H Z)^{-1} Z' P_H Y.$$

is biased and inconsistent.

Reduced form: $Y = \lambda(G(I - \lambda G)^{-1} X\beta) + X\beta + \eta,$ where

$$\eta = G(I - \lambda G)^{-1}\epsilon + \lambda BY + G(I - \lambda G)^{-1}\lambda B(I - \lambda G^*)^{-1}(X\beta + \epsilon).$$

Instrument exclusion restriction fails as

$$\text{Cov}(G(I - \lambda G)^{-1}X, \eta) \neq 0$$

Reduced form: $Y = \lambda(G(I - \lambda G)^{-1}X\beta) + X\beta + \eta$, where

$$\eta = G(I - \lambda G)^{-1}\epsilon + \lambda BY + G(I - \lambda G)^{-1}\lambda B(I - \lambda G^*)^{-1}(X\beta + \epsilon).$$

Instrument exclusion restriction fails as

$$\text{Cov}(G(I - \lambda G)^{-1}X, \eta) \neq 0$$

Solution

1. construct **corrected instruments** $G(I - \lambda(G + B))^{-1}X$ (in practice use expectation for units sampled incorrectly) – gets rid of third term in η
2. then left with **same problem as linear model** (unobserved BY also affected by instruments) – so rescale estimates as before.

How do we construct $E((GX)'BX)$ when $(G^*, B) \not\perp X$ without fitting a network formation model?

If we have a copula $C(x, d)$, we can sample

$$E(d_i|x_i) = \int_0^1 F_D^{-1}\left(\frac{\partial C(u_x, u_d; \theta)}{\partial u_x}\Big|_{u_x = F_X(x_i)}\right) dU_d.$$

Two step estimator.

1. Fit relevant copulas $C(F_X^{-1}, F_G^{-1}, \theta_1)$ to compute $\hat{BX}$.
2. Compute debiased estimator $\hat{\beta}$ given $\hat{BX}$.

Requires distributional assumptions (similar to Borusyak & Hull 2023), but we often have a good idea what the degree distribution should be (e.g Bacilieri et al. 2023, for production networks).

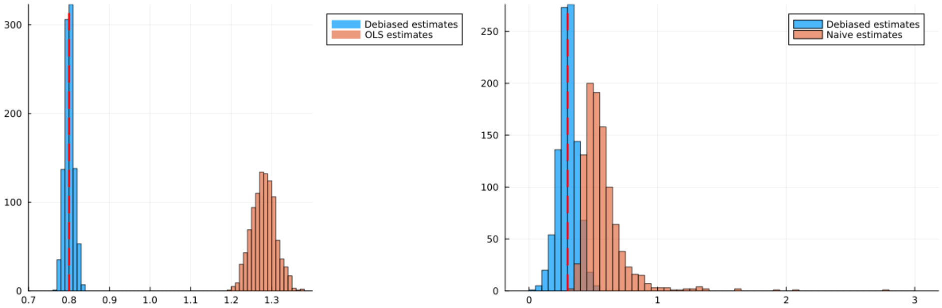
Simulations

Test size of bias and performance of new estimators on simulated networks sampled using common sampling rules.

$N = 1000$ agents with $d_i \sim U(0, 10)$ connected uniformly at random. Single binary treatment $X_i \in \{0.1\}$ distributed i.i.d across population (RCT on network) $X_i \sim \text{B}(0.3)$.

Sample networks using fixed choice design with $M = 5$ (undersampling, as in Add Health Datatset Harris 2009) and by assuming that each individual is connected to exactly 10 others (oversampling, as in Miguel & Kremer 2004).

Figure: Simulated spillover estimates from fixed choice design, M=5 (Add Health)

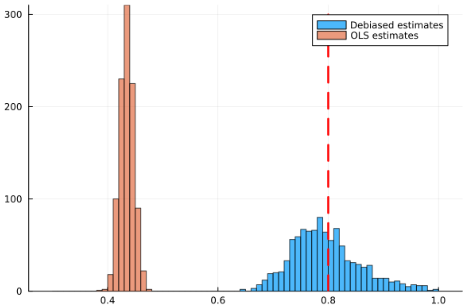


(a) OLS estimates (b) SAR estimates

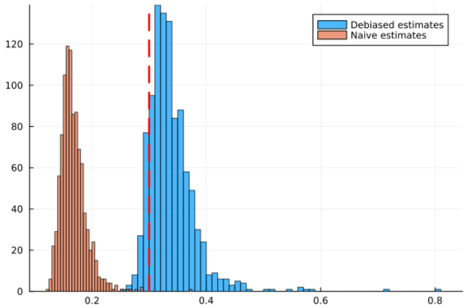
Notes: Red line denotes true parameter values of 0.8 and 0.3 respectively. $N = 1000$, $d_i \sim U(0, 10)$, $X_i \sim \mathrm{B}(0.3)$, $M = 5$.

Beginning Linear estimators

Figure: Spillover estimates from oversampled network (spatial spillovers)



(a) OLS estimates



(b) SAR estimates

Notes: Red line denotes true parameter values of 0.8 and 0.3 respectively. $N = 1000$, $d_i \sim U(0, 10)$, $X_i \sim \text{B}(0.3)$. Each individual has 10 sampled neighbours.

Application: production networks

Barrot & Sauvagnat (2016) study how effect of idiosyncratic shocks propagate in production networks running the regression

$$\Delta \text{SALES}_{it,t-4} = \alpha + \beta \text{ SUPPLIER\_HIT}_{it-4} + X_i \gamma + \epsilon_{it},$$

Use self-reported large suppliers of US public firms from Compustat – mean number of suppliers is 1.38, with a median of 0.000, evidence the dataset is undersampled (Herskovic et al. 2020).

$\text{SUPPLIER\_HIT}_{it-4}$ is a dummy that takes 1 if $GX > 0$ – can apply our results here.

Take d^G, and $p(SHOCK_{j,t-4} = 1)$ from their paper, construct $\bar{A}^{-1} = 0.07$
d^B from

1. more complete dataset covering similar firms (Factset)
2. estimated tail exponent of degree dist. adjusting for sampling from Herskovic et al. (2020), and
3. estimated tail exponent of degree dist. of complete (Belgian) production network from Bacilieri et al. (2023).

Table: Debiased spillover estimates

	Barrot & Sauvagnat (2016)	Factset	Herskovic et al. (2020)	Belgium
d^B	0	1.2	1.32	26.27
Estimate	-0.031	- 0.0159	-0.0151	-0.00160

Table: Mean missing links required to reject null of by significance level

	Reported	1%	5%	10%
Threshold	-0.031	-0.0225	-0.01764	-0.01476
d^B	0	0.474	0.953	1.39

Conclusions

Have shown that network over-and-under sampling biases conventional linear and nonlinear estimators for spillover effects and that the bias is large enough to matter in practice.

Have introduced debiased estimators based on the average number of unobserved links – requires one additional question in a survey

Have applied the estimators to construct unbiased estimates of the propagation of climate shocks in US firm-firm production network – sampling bias causes existing estimates to be 1.5-2 times larger than they should be.

References I

Aronow, P. M. & Samii, C. (2021), 'Estimating average causal effects under general interference, with application to a social network experiment', *Annals of Applied Statistics* **11**(4), 1912–1947.

Bacilieri, A., Borsos, A., Astudillo-Estevez & Lafond, F. (2023), 'Firm-level production networks: What do we (really) know?'.

Banerjee, A., Chandrasekhar, A., Duflo, E. & Jackson, M. (2013), 'The Diffusion of Microfinance', *Science* **341**(1236498), 363–341.

Barrot, J.-N. & Sauvagnat, J. (2016), ' Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks', *The Quarterly Journal of Economics* **131**(3), 1543–1592.

Beaman, L. A. (2011), 'Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S.', *The Review of Economic Studies* **79**(1), 128–161.

References II

Bloom, N., Schankerman, M. & Van Reenen, J. (2013), 'Identifying technology spillovers and product market rivalry', *Econometrica* **81**(4), 1347–1393.

Blume, L., Brock, W., Durlauf, S. & Jayaraman, R. (2015), 'Linear social interactions models', *Journal of Political Economy* **123**(2), 444–496.

Borusyak, K. & Hull, P. (2023), 'Nonrandom Exposure to Exogenous Shocks', *Econometrica* **91**(6), 2155–2185.

Breza, E., Chandrasekhar, A. G., McCormick, T. H. & Pan, M. (2020), 'Using aggregated relational data to feasibly identify network structure without network data', *American Economic Review* **110**(8), 2454–84.

Calvó-Armengol, A., Patacchini, E. & Zenou, Y. (2009), 'Peer effects and social networks in education', *The Review of Economic Studies* **76**(4), 1239–1267.

Carrell, S. E., Sacerdote, B. I. & West, J. E. (2013), 'From natural variation to optimal policy? the importance of endogenous peer group formation', *Econometrica* **81**(3), 855–882.

References III

Chandrasekhar, A. & Lewis, R. (2016), 'Econometrics of sampled networks', *Mimeo* .

Coleman, J., Katz, E. & Menzel, H. (1957), 'The diffusion of an innovation among physicians', *Sociometry* **20**(4), 253–270.

Conley, T. G. & Udry, C. R. (2010), 'Learning about a new technology: Pineapple in ghana', *American Economic Review* **100**(1), 35–69.

Foster, A. D. & Rosenzweig, M. R. (1995), 'Learning by doing and learning from others: Human capital and technical change in agriculture', *Journal of Political Economy* **103**(6), 1176–1209.

Griffith, A. (2022), 'Name your friends, but only five? the importance of censoring in peer effects estimates using social network data', *Journal of Labour Economics* **40**(4), 779–805.

Harris, K. M. (2009), 'The national longitudinal study of ad-olescent to adult health (add health), waves i and ii, 1994–1996.', *Carolina Population Center, University of North Carolina at Chapel Hill.* .

References IV

Herskovic, B., Kelly, B., Lustig, H. & Van Nieuwerburgh, S. (2020), 'Firm volatility in granular networks', *Journal of Political Economy* **128**(11), 4097–4162.

Herstad, E. I. (2023), 'Estimating peer effects and network formation models with missing links', *Mimeo* .

Hseih, C.-S., Hsu, Y.-C., Ko, S., Kovářík, J. & Logan, T. (2024), 'Non-representative sampled networks: Estimation of network structural properties by weighting'.

Jaffe, A. (1986), 'Technological opportunity and spillovers of research-and-development - evidence from firms patents, profits, and market value', *American Economic Review* **76**(5), 984–1001.

Kelejian, H. H. & Prucha, I. (1998), 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances', *The Journal of Real Estate Finance and Economics* **17**(1), 99–121.

Lewbel, A., Qu, X. & Tang, X. (2022), 'Estimating Social Network Models with Missing Links', *Mimeo* .

References V

Manski, C. F. (1993), ‘Identification of Endogenous Social Effects: The Reflection Problem’, *The Review of Economic Studies* **60**(3), 531–542.

Miguel, E. & Kremer, M. (2004), ‘Worms: Identifying impacts on education and health in the presence of treatment externalities’, *Econometrica* **72**(1), 159–217.

Munshi, K. (2003), ‘Networks in the modern economy: Mexican migrants in the u. s. labor market’, *The Quarterly Journal of Economics* **118**(2), 549–599.

Newman, M. (2010), *Networks*, Oxford University Press, Oxford.

Oster, E. & Thornton, R. (2012), ‘Determinants of technology adoption: Peer effects in menstrual cup take-up.’, *Journal of the European Economic Association* **10**(6), 1263–1293.

Rapoport, A. & Horvath, W. J. (1961), ‘A study of a large sociogram’, *Behavioral Science* **6**(4), 279–291.

Yauck, M. (2022), ‘On the estimation of peer effects for sampled networks’.

Zhang, L. (2023), ‘Spillovers of program benefits with missing network links’.

Example: 'fixed choice' design (Newman 2010, Coleman et al. 1957, Calvó-Armengol et al. 2009, Oster & Thornton 2012, Banerjee et al. 2013)

$$(GX)_i = \begin{cases} (G^*X)_i - (BX)_i & \text{if } d_i > m \\ (G^*X)_i & \text{if } d_i \leq m. \end{cases}$$

Consider $X \sim B(p)$ i.i.d across nodes (e.g randomised intervention) and individuals report each link with probability q (Griffith 2022). Then

$$E(GX)_i = \begin{cases} \frac{5}{d_i} \sum_j g^*_{ij} p & \text{if } \sum_j g^*_{ij} > 5, \\ \sum_j g^*_{ij} p & \text{if } \sum_j g^*_{ij} \leq 5 \end{cases}, \text{ and } E(BX)_i = \begin{cases} \frac{d_i - m}{d_i} \sum_j g^*_{ij} p & \text{if } \sum_j g^*_{ij} > 5, \\ 0 & \text{if } \sum_j g^*_{ij} \leq 5. \end{cases}$$

Therefore $E(BX), E((GX)'BX) > 0$.

Assumption (OLS assumptions)

Assume the following about our data generating process **??**

1. (Y, G^*, B, X) are independently but not identically distributed over i,
2. $E(\epsilon | G^*, X) = 0$
3. $E(G^* X_i) = \xi_i$, $V(G^* X_i) = r_i^2$, and $\lim \frac{\sum_{i=1}^N E(|G^* X_i - \xi_i|^{2+\delta})}{(\sum_{i=1}^N r_i^2)^{\frac{2+\delta}{2}}} = 0$ for some $\delta > 2$,
4. $E(BX_i) = \nu_i$, $V(BX_i) = s_i^2$, and $\lim \frac{\sum_{i=1}^N E(|BX_i - \nu_i|^{2+\delta})}{(\sum_{i=1}^N s_i^2)^{\frac{2+\delta}{2}}} = 0$ for some $\delta > 2$,
5. ϵ are idependent and not identically distributed over i such that for some $\delta > 0$ $E(|u_i^2|^{1+\delta}) < \infty$ with conditional variance matrix
$$E(\epsilon \epsilon' | (G^* - B)X) = \Omega$$
which is diagonal.
6. $\text{plim} \frac{1}{N}((G^* - B)X)' \epsilon \epsilon' ((G^* - B)X)$ exists, is finite, and is positive definite. Additionally, for some $\delta > 0$ $E(|\epsilon_i^2((G^* - B)X)_{ij}((G^* - B)X)_{ik}|^{1+\delta}) < \infty$ for all

Assume that

1. (Y, G^*, B, X) are independently but not identically distributed over i,
2. $E(\epsilon | G^*, X) = 0$
3. ϵ are independent and not identically distributed over i such that for some $\delta > 0$ $E(|u_i^2|^{1+\delta}) < \infty$ with conditional variance matrix

$$E(\epsilon\epsilon' | (G^* - B)X) = \Omega$$

   which is diagonal.

4.

$$\text{plim } N^{-1} Z' P_H Z = Q_{ZZ}$$
$$\text{plim } N^{-1} Z' P_H Z_B = Q_{ZB}$$
$$\text{plim } N^{-1} Z' P_H = Q_{ZH}$$

   which are each finite nonsingular.

5. $|\lambda| < \frac{1}{||G||}, \frac{1}{||G^*||}$ for any matrix norm $||.||$.