The London School of Economics and Political Science

Essays in Econometric Theory

Kamila Nowakowicz

A thesis submitted to the Department of Economics of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, January 2025.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 49712 words.

Statement of co-authored work

I confirm that Chapter 3 was jointly co-authored with Professor Tatiana Komarova. I contributed 80% of the theoretical work while Tatiana was fully responsible for simulations and application.

Statement of inclusion of previous work

I confirm that Chapter 3 builds on my thesis for MRes in Economics I undertook at the London School of Economics and Political Science. The chapter has been fully rewritten, the result has been generalised an extended.

Abstract

We design and analyse nonparametric techniques for understanding the structure of network data, inference on network statistics, and testing for shape constraints.

In the first chapter, I look at symmetric binary exchangeable networks. Any such network can be characterised by a distribution over characteristics of nodes and a linking (graphon) function which gives the probability of link between any two nodes. To learn about the network structure, I propose a nonparametric estimator of the linking function. I provide conditions under which the estimator is uniformly consistent and a numerical procedure for choosing a tuning parameter. My procedure makes minimal assumptions and allows for moderate sparsity levels.

In the second chapter, I propose a bootstrap procedure which allows for valid inference on network statistics. It uses my nonparametric linking function estimator from Chapter 1 to generate bootstrap networks with a similar dependence structure to the original network. I prove that the distribution of the bootstrap network is consistent for the distribution of the original network, and I provide conditions under which bootstrap consistently recovers distributions of a class of functions related to U-statistics. I find good performance in Monte Carlo simulations and apply my procedure to the data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013).

In the third chapter, we propose a test for whether a nonparametric regression mean satisfies a shape restriction that varies within the domain of the regressor (e.g. (inverted) U-shaped, S-shaped). Our procedure extends the methodology of Komarova and Hidalgo (2023) to the setting where the points at which the shape changes are unknown and must be estimated, and the shapes may only appear after controlling for covariates. We provide a generalised transformation which achieves the same asymptotic distribution but adds robustness to the test and credibility to the conclusions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Nonparametric linking probabilities estimator for exchangeable networks

Abstract

Network data is gaining popularity in economic applications, but using it for estimation and inference remains challenging due to the complicated dependence structure. Many economic networks take the form of symmetric binary exchangeable networks, and any such network can be characterised by a distribution over characteristics of nodes and a linking (graphon) function which gives the probability of link between any two nodes. We explore this representation which allows summarising the dependence in a more tractable way, and to learn about the network dependence structure, we propose a nonparametric estimator for the linking probabilities. We provide conditions under which the estimator is uniformly consistent and we give a numerical procedure for choosing a tuning parameter. Our procedure makes minimal assumptions and allows for moderate sparsity levels. In Chapter 2 we explore the use of our estimator for inference on networks.

1.1 Introduction

Network data is becoming more prevalent in economics, especially in fields such as development or labour economics. It allows for modelling e.g. the spread of information (Banerjee, Chandrasekhar, Duflo, and Jackson (2013)) or spillover effects (Carter, Laajaj, and Yang (2021)), but it comes with a challenge: because the observations within a network exhibit strong dependence, standard techniques for estimation and inference which rely on the assumptions of independence or weak dependence become invalid. There is a need to develop approaches which account for the network structure.

One way forward relies on assuming that the complicated dependence can be summarised using a simpler representation. One such representation exists for symmetric binary exchangeable networks. Symmetric means the relationship is not directed (as opposed to directed networks in which e.g. an influencer has many followers but does not follow them back); binary means there either is a connection or there is not, the strength of the connection is not weighted; and exchangeable means that all individuals come from the same distribution, the distribution of the network is invariant under finite permutations of individuals (an exception could be a network with a 'superstar': an individual who is fundamentally different from everyone else, there is always precisely one person like that in any network). Many economically-relevant networks, such as collaboration networks (e.g. between co-authors publishing in a given journal) or social networks (e.g. friendship relationships on facebook, Chetty et al. (2022) or connections between households in a village, Breza and Chandrasekhar (2019)) can be naturally represented in this framework.

Assuming we are working with a network of this form, we can represent its distribution using two objects: a distribution over characteristics of individuals (nodes) and a binary linking (graphon) function which takes the characteristics of any two nodes as inputs and outputs the probability with which they are linked. Our focus is providing an estimate of the linking probabilities for any pair of individuals observed in our sample.

We propose a nonparametric method to estimate the linking probabilities which takes advantage of the information provided by the set of observed connections. We start by borrowing a distance from Auerbach (2022) (see also Zhang, Levina, and Zhu (2017)), which can be intuitively summarised as: people with similar sets of friends are similar to each other. If we observe two people with similar sets of neighbours, it likely happened because the linking function gave similar probabilities for their links with other individuals. This way we can identify people similar to any person i. We can then determine what proportion of these 'counterfactuals' of i are linked to a person j, providing an estimate of the link probability between i and j. Similarly, by swapping the roles of i and j, we can find what proportion of individuals similar to j are linked to i. Assuming that links are symmetric, we take the average of these two estimates as an estimate of the link probability between i and j.

We provide conditions under which our estimator achieves uniform consistency. These are mostly concerned with the allowed level of sparsity and the smoothness of the linking function. As is common in real-life networks, we assume sparsity, which means that the linking probability grows at a slower rate than the number of individuals in a network. However, to allow for estimation, we need to put bounds on the allowed sparsity level: if the network was too sparse we would not be getting enough information to ensure convergence of our estimators. The level we assume is typical in this literature. We also assume that the estimated function allows us to find sufficiently many good counterfactuals for any individual we could observe. This assumption is closely related to the assumptions of "Bi-Lipschitz" or "Piecewise-Lipschitz" used by e.g. Zhang, Levina, and Zhu (2017), Auerbach (2022).

Our estimator takes a form of a nonparametric kernel estimator which requires the researcher to choose a tuning bandwidth parameter: we need to indicate how similar the potential counterfactuals of i need to be to be given a positive weight in the estimation procedure. One of our contributions is proposing a cross-validation procedure for choosing a bandwidth parameter for our estimator.

Our estimator has many potential uses. In Chapter 2 we show that it may be used as an input in a bootstrap procedure, allowing us to generate new networks with a structure similar to that of an observed network, which we can use for valid inference on network statistics. Another potential use, which we showcase in the application in Chapter 2 and plan to extend in future work, relies on the fact that the estimator can be seen as a proxy for the strength of connections between individuals in a network and it can be used directly in forming models of network interactions. This could help mitigate tha bias resulting from observing only binary information on links when it is thought that the observed links provide a noisy signal about the real determinant of behaviour: the strength of connections between individuals. We believe the estimator may prove to have many more potential applications.

In Section 1.2 we summarise the related literature. The setup of the model is described in Section 1.3, where we also provide a definition of our estimator. Section 1.4 includes the statement of our main result: the uniform consistency of the linking probabilities estimator in Theorem 1. Section 1.5 shows results of Monte Carlo simulations. Section 1.6 concludes. The appendix start with a list of all notation. Section 1.A includes all proofs.

1.2 Related literature

Our idea for the nonparametric linking probabilities estimator was inspired by Auerbach (2022), who provides a way of controlling for a network-dependent latent covariate in a partially linear regression setting. Our estimator has been previously proposed by Zeleneev (2020) (whose focus is different than ours and who does not analyse the theoretical properties of the estimator).

The closest paper to ours is Zhang, Levina, and Zhu (2017). They provide an estimator based on a very similar idea but using nearest neighbours instead of kernels and a different norm for the distance. The key distinction is that they do not directly model sparsity (instead they allow the linking function to come from a family of Piecewise-Lipschitz functions which becomes richer as the sample size increases) and their choice of a tuning parameter is motivated by the asymptotic rates (rather than done numerically from the sample).

One motivation for providing a kernel estimator is that we wish to use our linking probability estimator in conjunction with our bootstrap procedure developed in Chapter 2. Abadie and Imbens (2008) show that bootstrapping nearest neighbours estimators can lead to invalid inference. Our estimator can be seen as a smoothed-out version of a nearest-neighbour estimator, where observations that provide a worse fit are given a lower weight, which could lead to more stable behaviour when used with bootstrap. We begin to explore this idea in the application in Chapter 2 and we plan to extend it in future work.

Other approaches to estimating the linking function usually rely on imposing a parametric structure. One of the most popular parametric forms is the stochastic block models: it assumes that each node belongs to one of finitely many blocks, and the probability of a link between any two nodes is fully determined by the blocks the nodes belong to. There have been many procedures proposed for these kinds of models, e.g. Lei and Rinaldo (2015) analyse the statistical properties of spectral clustering algorithms for community detection in stochastic block models like in Rohe, Chatterjee, and Yu (2011), Amini, Chen, Bickel, and Levina (2013) provide a pseudo-likelihood procedure for community detection and Guédon and Vershynin (2016) proposed a procedure based on Grothendieck's inequality.

A recent addition to this literature is Kitamura and Laage (2024) who develop a method of incorporating observed covariates (in addition to an adjacency matrix) when estimating stochastic block models. This extension has a lot of economic relevance, as in many applications we do have access to observed covariates and we should take advantage of this information in estimating our model. Using a similar framework with our procedure would be an interesting extension of our model.

Other parametric frameworks include the assumption of a dot product linking function as in Levin and Levina (2019) see Athreya et al. (2018) for a survey of work on estimation and statistical properties.

1.3 Model: setup and definitions

1.3.1 Setup

We follow the standard setup in the literature known as the latent space model.

We observe an adjacency matrix A which corresponds to an undirected, unweighted graph on n nodes (also referred to as individuals) indexed by $i \in \{1, 2, \dots, n\}$. The matrix is symmetric, has zeros on the main diagonal and ones in positions corresponding to edges in the graph ($A_{ij} = 1$ if and only if there is an edge between nodes i and j). Each node i is characterised by a vector of unobserved features[1] ξ_i, drawn independently from their common distribution F_0 with support $Supp(\xi_i)$. We denote the vector of all $\{\xi_i\}_{i=1}^n$ by ξ. We assume that the distribution has no point mass,[2] i.e. for $\xi_i, \xi_j \sim F_0$ we have $P_{F_0}(\xi_i = \xi_j) = 0$. We impose more assumptions[3] on F_0 in Assumption 1.2.

Let $h_{0,n} : Supp(\xi_i) \times Supp(\xi_i) \to [0, 1]$ be a symmetric, measurable linking function[4] which can be decomposed as:

$$h_{0,n}(u, v) = \rho_n w_0(u, v) \tag{1.1}$$

where $\int w_0(u, v) dF_0(u) dF_0(v) = 1$.

For each pair of nodes i, j, $h_{0,n}(\xi_i, \xi_j)$ maps their unobserved characteristics ξ_i, ξ_j into the probability of a link (edge) between them, i.e. the probability with which $A_{ij} = 1$. We treat the linking function as unknown, making minimal assumptions on its properties in Assumption 1.2: we require that for each input there is a neighbourhood of sufficiently large measure in which the behaviour of the function remains similar. Importantly, we do not require a specific form (e.g. random dot product structure: $h_{0,n}(\xi_i, \xi_j) = \xi_i' \xi_j$ like in Levin and Levina (2019)), we do not impose any shape constraints (e.g. that the function is strictly increasing in its inputs).

The decomposition into ρ_n and w_0 can be seen as a normalisation which allows us to interpret ρ_n as the expected edge density (the marginal probability of an edge between two nodes). We assume $\rho_n \to 0$ as $n \to \infty$, which captures the common feature of real economic networks known as sparsity. Intuitively, it says that the number of expected friends grows at a slower

[1] This corresponds to the vector of latent positions X_i in Levin and Levina (2019).

[2] This is without loss of generality: if we had a distribution with a point mass we could define a new support of ξ and a new F_0 in which the point mass would be replaced by a region of ξ of total measure equal to the probability at the original point.

[3] The assumptions are implicit and would be implied by F_0 bounded above and separated away from zero with $h_{0,n}$ piecewise Lipschitz.

[4] The linking function has been referred to as the coupling function $g(.,.)$ in Zeleneev (2020) and the graphon function in Green and Shalizi (2022).

rate than the size of the network: no matter how large the potential pool of friends is, people tend to have a fairly small friendship group. This causes issues for estimation because, even as the size of the network grows at a rate n, the amount of information about the links of a specific node i grows at a slower rate of $\rho_n n$. In the extreme case of $\rho_n n$ being bounded we cannot hope to get consistency of our estimates. In our results we specify bounds on the rate at which ρ_n approaches zero which still allow us to reliably estimate parameters and their distributions. For the linking probabilities estimator we require that the density decreases at a slower rate than $\sqrt{\frac{\log(n)}{n}}$ (see Assumption 1.1).

w_0 is the underlying linking/graphon function after accounting for sparsity. While w_0 cannot be interpreted directly as a probability, it has similar properties, e.g. it is bounded.[5] This is the function which determines the data generating process and the function the statistics of which we want to analyse. Although in a sample of size n we encounter its rescaled version $h_{0,n}$, for any asymptotic results we need to remove the effect of sparsity and we look at normalisations which are function of $\frac{h_{0,n}}{\rho_n}$.

To capture the way in which the linking function $h_{0,n}$ is translated into the observed links in A we introduce a random noise parameter: for $1 \leq i \leq j \leq n$ let $\eta_{ij} \overset{ind}{\sim} \mathcal{U}[0,1]$ be independent of ξ. We denote the vector of η_{ij} by η. We assume:[6]

$$\begin{aligned} A_{ij} &= A_{ji} = \mathbb{1}\left(h_{0,n}(\xi_i, \xi_j) \geq \eta_{ij}\right) \\ A_{ii} &= 0. \end{aligned} \tag{1.2}$$

Note that $E(A_{ij}|\xi_i, \xi_j) = P(A_{ij} = 1|\xi_i, \xi_j) = h_{0,n}(\xi_i, \xi_j) = \rho_n w_0(\xi_i, \xi_j)$. To distinguish between adjacency matrices based on the true and estimated/simulated inputs we sometimes explicitly write A as a function: $A(h_{0,n}(\xi), \eta)$.

Our goal is to find a good estimator $\hat{h}_n$ of the linking function $h_{0,n}$ at all observed pairs ξ_i, ξ_j.

Remark. *We treat the model as the true data generating process, but our results could be considered more general: by the Aldous-Hoover Theorem (originally proven independently in Aldous (1981) and Hoover (1979), see also discussion in Kallenberg (1989) and Orbanz and Roy (2014)), as long as the distribution of the infinite array $X = (X_{ij},\ i, j \in \mathbb{N})$ of random variables*

[5] This is a common assumption in the literature, though it is sometimes relaxed to allow $w_0(u,v) \in \mathbb{R}_+$ and let $h_{0,n}(u,v) = \min\{w_0(u,v), 1\}$. This affects the interpretation of ρ_n as the density and makes it more difficult to infer $h_{0,m}$ from $h_{0,n}$. Our results could be generalised to allow for unbounded w_0 at the expense of more complicated proofs and additional assumptions on bounded moments of w_0 or its functions.

[6] This is one specific way of achieving:

$$\begin{aligned} A_{ij}|\xi = A_{ji}|\xi &\overset{ind}{\sim} \text{Bernoulli}\left(h_{0,n}(\xi_i, \xi_j)\right) \\ A_{ii} &= 0 \end{aligned}$$

(corresponds to the limit of our A for a fixed level of sparsity ρ_n) is invariant under joint permutations of the rows and columns (i.e. if (p_i) is a permutation of indices, the distribution of (X_{ij}) is the same as that of $\left(X_{p_i,p_j}\right)$: the distribution is invariant to relabelling of the individuals) there exist i.i.d. random variables α, ξ_i, η_{ij}, $i,j \in \mathbb{N}$ and a measurable function f for which:

$$X_{ij} = f(\alpha, \xi_i, \xi_j, \eta_{ij}), \quad i,j \in \mathbb{N}.$$

If, as in our case, X_{ij} are binary and symmetric, the representation simplifies[7] *to:*

$$\begin{aligned} X_{ij} &= \mathbb{1}\left(w(\alpha, \xi_i, \xi_j) \geq \eta_{ij}\right), \quad i,j \in \mathbb{N}, i \neq j; \\ X_{ii} &= 0. \end{aligned}$$

We treat the underlying binary jointly exchangeable infinite array as a mixture of processes of the form in Eq. (1.2). To obtain the representation in Eq. (1.2), we condition on the realised value of α.

1.3.2 Distance: definition and estimator

Based on the observed matrix A, we want to estimate the linking probability for any pair of nodes. We start by defining a distance between individuals i and j, taking the measure[8] from Auerbach (2022). Intuitively, if two people have similar friendship groups, they should be similar to each other: they likely ended up with similar friendship groups because their linking functions were similar. We let $\varphi(\xi_i, \xi_t) = E\left(w_0\left(\xi_i, \xi_s\right) w_0\left(\xi_t, \xi_s\right) \middle| \xi_i, \xi_t\right) = E\left(\frac{A_{is}}{\rho_n}\frac{A_{ts}}{\rho_n}\middle| \xi_i, \xi_t\right)$ be a function measuring the probability of a common friend between i and t, normalised to remove the effect of sparsity. Similarly, $\varphi(\xi_j, \xi_t)$ gives a normalised measure of the probability of common friends between j and t. To measure the similarity in friendship groups between i and j we look at the expected difference $\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t)$ for any individual t. This motivates

[7] We can set

$$w(\alpha, x, y) = \int_0^1 f(\alpha, x, y, \eta) d\eta;$$

see Corollary III.6. in Orbanz and Roy (2014).

[8] Auerbach (2022) refers to $\varphi_{\xi_i}(\tau) = \varphi(\xi_i, \tau)$ as the codegree function of agent i. In his model there is no sparsity: $\rho_n = 1$, which is why he does not need the normalisation by $\frac{1}{\rho_n}$.

the definition of distance between i and j:

$$
\begin{aligned}
d_{ij} &= \sqrt{E\left(\left(\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t)\right)^2 \middle| \xi_i, \xi_j \right)} && (1.3)\\
&= \sqrt{E\left(\left(E\left(w_0\left(\xi_t, \xi_s\right)\left(w_0\left(\xi_i, \xi_s\right) - w_0\left(\xi_j, \xi_s\right)\right) \middle| \xi_i, \xi_j, \xi_t\right)\right)^2 \middle| \xi_i, \xi_j \right)} && (1.4)\\
&= \sqrt{E\left(E\left(\frac{A_{ts}}{\rho_n}\left(\frac{A_{is}}{\rho_n} - \frac{A_{js}}{\rho_n}\right) \middle| \xi_i, \xi_j, \xi_t \right)^2 \middle| \xi_i, \xi_j \right)}. && (1.5)
\end{aligned}
$$

After an appropriate normalisation by the sparsity level ρ_n, we get an expression in terms of the linking function $h_{0,n} = \rho_n w_0$ at sample size n:

$$
\begin{aligned}
\rho_n^2 d_{ij} &= \sqrt{E\left(\left(E\left(h_{0,n}\left(\xi_t, \xi_s\right)\left(h_{0,n}\left(\xi_i, \xi_s\right) - h_{0,n}\left(\xi_j, \xi_s\right)\right) \middle| \xi_i, \xi_j, \xi_t\right)\right)^2 \middle| \xi_i, \xi_j \right)} && (1.6)\\
&= \sqrt{E\left(E\left(A_{ts}\left(A_{is} - A_{js}\right) \middle| \xi_i, \xi_j, \xi_t\right)^2 \middle| \xi_i, \xi_j \right)}. && (1.7)
\end{aligned}
$$

Eq. (1.6) highlights the close relation between the normalised distance and the similarity between the linking functions of i and j at sample size n: a low value of $\rho_n^2 d_{ij}$ means i and j are similar to each other in the sense that their $h_{0,n}(\xi_i, \cdot)$ and $h_{0,n}(\xi_j, \cdot)$ are close. We exploit this when defining an estimator for $h_{0,n}$. The normalised expression is also attractive because the sample equivalent of its representation in Eq. (1.7) provides us with an estimate of $\rho_n^2 d_{ij}$:

$$
\rho_n^2 \hat{d}_{ij} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\sum_{s=1}^{n} A_{ts}\left(A_{is} - A_{js}\right)\right)^2}. \qquad (1.8)
$$

Remark. *If we needed to estimate $\hat{d}_{ij}$ without the normalisation we could substitute the estimated density:*

$$
\hat{\rho}_n = \frac{1}{\binom{n}{2}} \sum_{i \leq i < j \leq n} A_{ij} \qquad (1.9)
$$

for the unknown ρ_n. However, in practice the way we use the distance is with a normalisation by a bandwidth parameter a_n (chosen by the researcher), we look at functions of: $\frac{\rho_n^4 \hat{d}_{ij}^2}{a_n} \equiv \frac{\hat{d}_{ij}^2}{b_n}$, and we can think of the ρ_n as being absorbed into the renormalised bandwidth b_n.

Remark. *We could also consider estimators of related distances from Zeleneev (2020), who took it from Zhang, Levina, and Zhu (2017): $\rho_n \hat{d}_{ij}^{(\infty)} = \left(\max_{t \neq i,j} \left|\frac{1}{n}\sum_{s \neq i,j,t} A_{ts}(A_{is} - A_{js})\right|\right)^{\frac{1}{2}}$ and from Lovász (2012) (sections 13.4, 15.4): $\rho_n^2 \hat{d}_{ij}^{(1)} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{1}{n}\sum_{s \neq i,j,t} A_{ts}(A_{is} - A_{js})\right|$. All of these distances are based on the same idea but average $\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t)$ using L_2, L_∞ and L_1 distances, respectively.*

Alternative distances could be used as well. However, note that the direct comparison of the *friendship groups does not provide a good distance.*[9]

The distance we use looks at the number of second-order connections (friends of friends). We could consider comparing higher order connections, but this would come at a cost of more difficult computation and it would require stronger assumptions on the allowed sparsity level (the reason why we need ρ_n to go to zero slower than $\sqrt{\frac{\log(n)}{n}}$ rather than $\frac{\log(n)}{n}$ is closely related to the fact that our estimate of distance requires a normalisation by ρ_n^2).

1.3.3 Linking probabilities estimator

We are now ready to define an estimator $\hat{h}_n$ for the linking function $h_{0,n}$ at all pairs of realised ξ_i, ξ_j. We sometimes refer to this object as the 'linking function estimator,' since it provides an approximation to the true linking function, but note that we provide the approximation only at specific points and we do not characterise the estimator as a function of its latent inputs ξ_i, ξ_j. Hence we usually call it the 'linking probabilities estimator,' since it is estimating the probabilities of links between all pairs of observed individuals. We rely on a kernel approximation: let $K(\cdot)$ be a kernel function (for properties see Assumption 1.3), let a_n be a bandwidth parameter (for its rates of convergence see Assumption 1.4, see Section 1.3.4 for a method of choosing a bandwidth). We can estimate $h_{0,n}(\xi_i, \xi_j)$ as:

$$\hat{h}_n(\xi_i, \xi_j) = \frac{\tilde{h}_n(\xi_i, \xi_j) + \tilde{h}_n(\xi_j, \xi_i)}{2} \tag{1.10}$$

where

$$\tilde{h}_n(\xi_i, \xi_j) = \frac{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) A_{tj}}{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right)}. \tag{1.11}$$

$\tilde{h}_n(\xi_i, \xi_j)$ is a local weighted average which puts the highest weights on the individuals most

[9] This could be defined as:

$$\rho_n d_{ik}^s = \sqrt{E_l\left[(A_{li} - A_{lk})^2 \middle| \xi_i, \xi_k\right]}, \quad \rho_n \hat{d}_{ik}^s = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(A_{li} - A_{lk})^2}$$

Suppose $h_{0,n}(\xi_i, \xi_j) = \frac{1}{2}$ for all ξ_i, ξ_j. Since all individuals share te same linking probabilities, we would like their distance to be zero. But i and k are expected to have around $\frac{n}{2}$ friends each, with only $\frac{n}{4}$ overlapping.

$$\sqrt{\frac{1}{n}\sum_{l=1}^{n}(A_{li} - A_{lk})^2} = \sqrt{\frac{1}{n}\sum_{l=1}^{n} A_{li}^2 + A_{lk}^2 - 2A_{li}A_{lk}} = \sqrt{\frac{1}{n}\sum_{l=1}^{n} A_{li} + \frac{1}{n}\sum_{l=1}^{n} A_{lk} - \frac{2}{n}\sum_{l=1}^{n} A_{li}A_{lk}}$$

$$\xrightarrow{p} \sqrt{E[A_{li}] + E[A_{lk}] - 2E[A_{li}A_{lk}]} = \sqrt{\frac{1}{2} + \frac{1}{2} - 2\frac{1}{4}} = \frac{1}{4} \not\to 0$$

Since the expected size of i's and k's common neighbourhood with any l is around $\frac{n}{4}$, our original distance is close to 0.

similar to i. The bandwidth a_n controls the required level of similarity beyond which we use zero weights. Each person t with $\hat{d}_{it}$ sufficiently close to zero can be seen as a counterfactual to i, someone with a very similar linking function (i.e. a small $h_{0,n}(\xi_i, \cdot) - h_{0,n}(\xi_t, \cdot)$). The proportion of people similar to i who are linked to j gives an estimate of the link probability between i and j. The observation with $t = j$ is excluded because we assume there are no self-link: $A_{jj} = 0$ is not defined in terms of $h_{0,n}$. Adding a non-zero weight on this observation would introduce bias.

To get $\hat{h}_n(\xi_i, \xi_j)$ we take advantage of the symmetry of links, we repeat the estimation swapping the roles of i and j and take an average of the two estimates.

We show that this estimator is uniformly consistent for $h_{0,n}$ in Theorem 1.

Remark. *The estimator $\hat{h}_n$ has been proposed, but not analysed by Zeleneev (2020). We could also use a related estimator:*

$$\hat{h}_n^{(K2)}(\xi_i, \xi_j) = \frac{\sum_{\substack{t=1\\t\neq s}}^n \sum_{s=1}^n K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) K\left(\frac{\rho_n^4 \hat{d}_{js}^2}{a_n}\right) A_{ts}}{\sum_{\substack{t=1\\t\neq s}}^n \sum_{s=1}^n K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) K\left(\frac{\rho_n^4 \hat{d}_{js}^2}{a_n}\right)}.$$

Zhang, Levina, and Zhu (2017) propose a closely related estimator which uses the nearest-neighbour idea:

$$\hat{h}_n^{(NN1)}(\xi_i, \xi_j) = \frac{\tilde{h}_n^{(NN1)}(\xi_i, \xi_j) + \tilde{h}_n^{(NN1)}(\xi_j, \xi_i)}{2} \text{ where } \tilde{h}_n^{(NN1)}(\xi_i, \xi_j) = \frac{\sum_{t\in\mathcal{N}_i} A_{tj}}{\|\mathcal{N}_i\|}$$

where $\mathcal{N}_i$ denotes the set of neighbours of i. They show that the optimal size of the neighbourhood, $\|\mathcal{N}_i\|$, should grow at the rate of $(n \ln(n))^{1/2}$. We could also use a nearest-neighbour approach in both inputs simultaneously (again, mentioned, and this time analysed, by Zeleneev (2020)):

$$\hat{h}_n^{(NN2)}(\xi_i, \xi_j) = \frac{\sum_{t\in\mathcal{N}_i} \sum_{s\in\mathcal{N}_j} A_{ts}}{\|\mathcal{N}_i\| \|\mathcal{N}_j\|}.$$

1.3.4 Choice of bandwidth

The linking probabilities estimator relies on a bandwidth parameter chosen by the researcher. We propose a cross-validation procedure which allows choosing the bandwidth in an automated way from the observed sample.

The idea is to choose a bandwidth for which $\hat{h}_n$ best explains the observed network A, *if we leave out A_{ij} when estimating A_{ij}*. The reason for leaving out A_{ij} is that if we do not, we are trying to estimate A_{ij} using a set of observations which include A_{ij}, hence we can estimate it perfectly. We just need to choose $a_n \simeq 0$, this puts weight one on A_{ij} and zero on all other observations, leading to a perfect prediction of A but a poor choice of bandwidth. This issue of overfitting can be avoided by removing the observation A_{ij} from the model predicting A_{ij}.

We firstly define a leave-one-out version of $\hat{h}_n$:

$$\begin{aligned}
\tilde{h}_n^-(\xi_i, \xi_j) &= \frac{\sum_{\substack{t=1\\t\neq i,j}}^n K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) A_{tj}}{\sum_{\substack{t=1\\t\neq i,j}}^n K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right)} \\
\hat{h}_n^-(\xi_i, \xi_j) &= \frac{\tilde{h}_n^-(\xi_i, \xi_j) + \tilde{h}_n^-(\xi_j, \xi_i)}{2}.
\end{aligned}$$

and then use it to obtain an estimate for the log-likelihood:

$$\ell(A, a_n) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} \log\left(\hat{h}_n^-(\xi_i, \xi_j)\right) + (1 - A_{ij}) \log\left(1 - \hat{h}_n^-(\xi_i, \xi_j)\right). \tag{1.12}$$

We choose a_n which maximises the above expression to be our bandwidth:

$$\hat{a} = \max_{a_n} \ell(A, a_n). \tag{1.13}$$

1.4 Main result

In this section we state our main result which characterises the conditions under which the linking probabilities estimator is consistent.

1.4.1 Consistency of the linking probabilities estimator

We start by listing our assumptions.

Assumption 1 (The Assumptions for Uniform Consistency of the Linking Function Estimator)**.** *We make the following assumptions:*

1.1 $\frac{1}{\rho_n} = o\left(\sqrt{\frac{n}{\log(n)}}\right)$.

1.2 Let $N(\xi_j, \delta) = \left\{\xi_k : \sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta\right\}$ *denote the neighbourhood of* ξ_j *of size* δ *and let* $\omega(\delta) = \inf_{\xi_j \in Supp(\xi_j)} P\left(\xi_k \in N(\xi_j, \delta) | \, \xi_j\right)$*. There exist some* $\alpha, C > 0$ *such that* $\omega(\delta) \geq \left(\frac{\delta}{C}\right)^{\frac{1}{\alpha}}$ *for all* $\delta > 0$.

1.3 $K(\cdot)$ *is a kernel function which is*

- *a continuous bounded probability density function (non-negative:* $K(u) \geq 0$*, integrates to 1:* $\int K(u)du = 1$*),*

- *non-zero on a bounded support: there exists a $D \in \mathbb{R}$ such that $\forall |u| > D : K(u) = 0$,*
- *positive close to 0: there exist positive constants C_1, C_2 such that $K(u) \geq C_1$ whenever $|u| \leq C_2$,*
- *Lipschitz continuous: there exists $C > 0$ such that $|K(u) - K(v)| \leq C|u - v|$.*

1.4 The bandwidth can be written as $a_n = \rho_n^4 b_n$ for some $b_n = o(1)$ and $\frac{1}{b_n} = o\left(\left(\frac{n\rho_n^2}{\log(n)}\right)^{\frac{\alpha}{1+2\alpha}}\right)$.

We now discuss the assumption and provide intuition.

Assumption 1.1 is our sparsity assumption. It gives a lower bound on how sparse a model can be for our estimator to remain consistent. Intuitively, the only informative observations are the links, and their number grows at a slower rate than the sample size: we expect on average $n\rho_n$ links in a sample of n individuals. Our model works well if the number of links increases at a rate faster than $\sqrt{n \log(n)}$. This is analogous to the assumption in Zhang, Levina, and Zhu (2017), with the exception that they model the increasing difficulty in estimation with n by allowing $\delta(n) \to 0$ instead of having a sparsity parameter $\rho_n \to 0$.

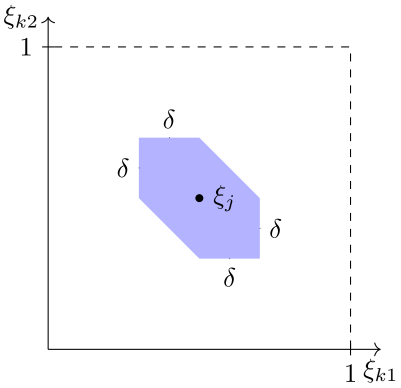
Assumption 1.2 ensures that the neighbourhoods for all observations are sufficiently large. We can think of it as a "continuity" condition for w_0, analogous to that assumed by Auerbach (2022): for all $\delta > 0$:

$$\inf_{\xi_j \in Supp(\xi_j)} P_{\xi_k \sim F_0}\left(\sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta \middle| \xi_j\right) \geq \left(\frac{\delta}{C}\right)^{\frac{1}{\alpha}}.$$

i.e., for each $\xi_j \in Supp(\xi_j)$ there exists a sufficiently large positive measure of ξ_k with very similar friendship groups: such that $|w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta$ holds for all ξ_t. The consequences of this assumption are similar to those of the Piecewise-Lipschitz assumption in Definition 2 of Zhang, Levina, and Zhu (2017) (in the proof of Theorem 2.4.1 in Chapter 2 we show that under Assumption 1.2 $\forall \varepsilon > 0$ $\exists K < \infty$ such that $Supp(\xi_i)$ can be split into K disjoint regions, each of size at least $\left(\frac{\varepsilon}{C}\right)^{\frac{1}{\alpha}}$, such that for any two points u, v which fall in the same region we have $\sup_{\xi_t} |w_0(\xi_t, u) - w_0(\xi_t, v)| \leq 2\varepsilon$ and there exists a set of points a_k, each from a different region, such that if $k \neq j$ we have $\sup_{\xi_t} |w_0(\xi_t, a_k) - w_0(\xi_t, a_j)| > \varepsilon$).

The example below shows that we can also think of this assumption as ensuring that the distribution of ξ_i is bounded away from zero while w_0 is sufficiently smooth:

Example. *For an example of Assumption 1.2 and some intuition on what α means, suppose that $\xi_i \in [0,1]^2$ and $w_0(\xi_i, \xi_j) = 2\xi_i \cdot \xi_j$. ξ_k satisfies $\sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta$ if it falls within a region in $[0,1]^2$ centred at ξ_j with a radius proportional to δ and area proportional to δ^2. If ξ_j is at least δ away from all boundaries of the support, the region has the following shape:*



and area $3\delta^2$. If the point ξ_j is closer to the boundary, we get the subset of this region which overlaps with $[0,1]^2$. The smallest area possible is $\frac{\delta^2}{2}$. If the distribution of ξ_i is uniformly bounded away from zero, the measure of ξ_j that satisfy the condition is at least proportional to δ^2, hence $\omega(\delta) \geq \left(\frac{\delta}{C}\right)^2$. A sufficient condition for our assumption is if the distribution of ξ_i is uniformly bounded away from zero and w_0 is piecewise Lipschitz. If w_0 is a well-behaved function and Supp$(\xi_i) \subset \mathbb{R}^d$, we would expect $\frac{1}{\alpha} = d$. We can think of $\frac{1}{\alpha}$ as a measure of complexity of the feature space: the more complex ξ_i are the harder the estimation.

Assumption 1.3 gives a list of fairly standard assumptions on the form of the kernel function. These are not restrictive as the kernel is chosen by the researcher and many of the standard kernels (e.g. the Epanechnikov kernel: $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}(|u| < 1)$ or the triangular kernel: $K(u) = (1-|u|)\mathbb{1}(|u| < 1)$) satisfy all the requirements.

Assumption 1.4 specifies the range of bandwidths for which we can guarantee the correct asymptotic behaviour. The bandwidth a_n is a product of ρ_n^4, which cancels out the normalisation in $\left(\rho_n^2 \hat{d}_{it}\right)^2$, and $b_n \to 0$ which ensures that $\frac{\rho_n^4 \hat{d}_{it}^2}{a_n} = \frac{\hat{d}_{it}^2}{b_n} \to \infty$ for all $i \neq t$. ρ_n can be estimated and b_n is chosen by the researcher. As the effective dimension of the support of ξ_i increases, i.e. α decreases, the estimation becomes more difficult and we need b_n to go to zero at a slower rate.

Theorem 1. *Under Assumption 1:*

$$\max_{i,j} \left| \frac{\hat{h}_n(\xi_i, \xi_j) - h_{0,n}(\xi_i, \xi_j)}{\rho_n} \right| \xrightarrow{p} 0.$$

Remark. *Notice that we can write $h_{0,n}(\xi_i, \xi_j) = \rho_n w_0(\xi_i, \xi_j)$, decomposing the linking function into a bounded function w_0 which does not depend on n and the sparsity $\rho_n \to 0$. Without the normalisation by $\frac{1}{\rho_n}$, the difference $\hat{h}_n - h_{0,n}$ would trivially go to zero because both components go to zero at the rate ρ_n. In the statement of Theorem 1 we normalise by $\frac{1}{\rho_n}$ to show that, even after removing the trend to zero, the estimate of the linking function approaches its true value.*

Remark. *Note on notation: we use* $\max_{i,j}$ *to refer to maximising over indices in a specific sample of size* n*: it is a shorthand notation for* $\max_{i,j\in\{1,2,\ldots,n\}}$*. We later use* $\max_{\xi_i}$ *which refers to maximising over all* $\xi_i \in Supp(\xi_i)$*, i.e. all possible values in the support, not the set of realised values in a specific sample.*

Remark. *Theorem 1 shows uniform convergence of* $\hat{h}_n(\xi_i, \xi_j)$ *to* $h(\xi_i, \xi_j)$*, where "uniform" refers to convergence over all pairs of nodes in the original graph. The reason why we do not look at uniformity*[10] *over all general points in the underlying sample space of* ξ_i *is because our estimator of distance* $\hat{d}_{ij}$ *is defined in terms of similarity of friendship groups, hence it can only be estimated for one of the observed individuals. In our procedure we never estimate* ξ_i *directly, we do not put strong assumptions on the space it comes from, and we do not have a way of estimating* $\hat{d}$*, and hence* $\hat{h}_n$*, at a general point* (u, v) *outside of our realised set of observed individuals.*

However, the results we show later in Theorem 2.4.1 can be seen as an extension of Theorem 1 to the whole support of ξ_i*: under the assumption Assumption 1.2, for any* $\xi_i \in Supp(\xi_i)$*, if* n *is high enough, with high probability we can observer* ξ_j *similar enough to* ξ_i *that* $\hat{h}_n$ *evaluated at* ξ_j *provides a good approximation to* ξ_i *and the frequencies with which we observe different values of* $\hat{h}_n$ *is representative of the frequencies of similar values of the true* $h_{0,n}$ *over the support of* ξ_i*.*

One may be interested in using an alternative notion of distance. In the result below we the characterise conditions a distance needs to satisfy to get the conclusions of Theorem 1.

Lemma 1. *Let* $\mathscr{d}_{ij}$ *be a distance between* ξ_i, ξ_j *such that there exist constants* $C_1, C_2 < \infty$*,* $\beta > 0$*,* $\gamma > 0$*,* $\mu > 1$ *for which* $\forall \xi_i, \xi_t \in Supp(\xi_i)$*:*

$$C_1 \mathscr{d}_{it}^{\beta} \leq \sup_{\xi_j} \left| w_0\left(\xi_t, \xi_j\right) - w_0\left(\xi_i, \xi_j\right) \right| \leq C_2 \mathscr{d}_{it}^{\gamma}$$

and let $\hat{\mathscr{d}}_{ij}$ *be a consistent estimate of* $\mathscr{d}_{ij}$ *and let* $\mu \in \mathbb{R}$ *be the smallest constant*[11] *for which*

$$\rho_n^{\mu} \max_{i,j} \left| \hat{\mathscr{d}}_{ij} - \mathscr{d}_{ij} \right| \xrightarrow{p} 0.$$

Under Assumption 1 with Assumption 1.1 replaced with $\frac{1}{\rho_n} = o\left(\left(\frac{n}{log(n)}\right)^{\frac{1}{\mu}}\right)$ *and the final condition in Assumption 1.4 replaced with* $\frac{1}{b_n} = o\left(\left(\frac{n\rho_n^{\mu}}{\log(n)}\right)^{\frac{\alpha}{2\beta+2\alpha}}\right)$ *we get the conclusions of Theorem 1 for* $\hat{h}_n$ *based on* $\hat{\mathscr{d}}_{ij}$*.*

[10] For example, if $\xi_i \sim \mathcal{U}[0,1]$ we could be interested in showing $sup_{u,v\in[0,1]} \left|\frac{\hat{h}_n(u,v) - h_{0,n}(u,v)}{\rho_n}\right| \xrightarrow{p} 0$.

[11] In our case the normalisation is $\mu = 2$ because we are looking at second order friendship groups: comparing number of common friends. If we looked at higher order statistics (e.g. numbers of friends of friends of friends) we would need a higher normalisation and we would get a stronger restriction on allowed sparsity level.

The main change is to the allowed level of sparsity: the harder the estimate of distance is to obtain, the stronger assumptions on sparsity we need to impose. The change in the assumptions on the rate of convergence of the bandwidth b_n is not important as this value is chosen by the researcher.

1.5 Simulations

We test the performance of our procedure by simulating a number of networks with known linking functions and checking how well our method can recover the probabilities of links. We consider ξ_i uniformly distributed between 0 and 1 and the following linking functions:

1. dot product function: $h(\xi_i, \xi_j) = \rho_n \xi_i \xi_j$. This is the parametric form assumed by Levin and Levina (2019), it is a relatively simple function and a good benchmark.

2. horseshoe function: $h(\xi_i, \xi_j) = \frac{\rho_n}{2}\left(e^{-200\left(\xi_i - \xi_j^2\right)^2} + e^{-200\left(\xi_j - \xi_i^2\right)^2}\right)$. This function was also used by Green and Shalizi (2022) and Wang (2016), who described it as “a challenging example for graphon estimation.”

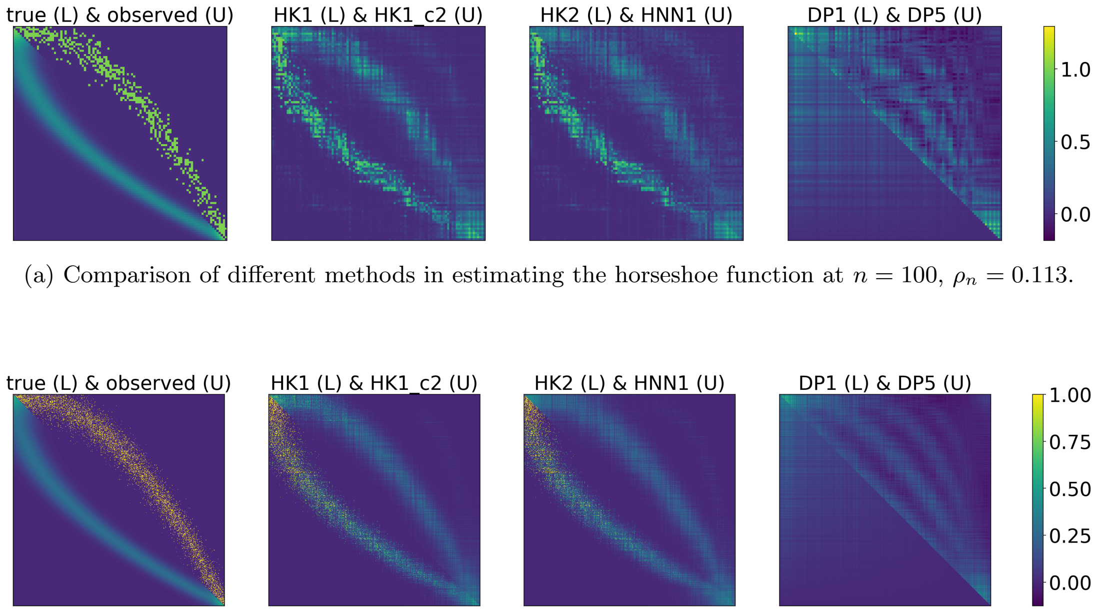
3. high-density function:
$h(\xi_i, \xi_j) = \frac{\rho_n}{0.975}\left(1 - \mathbb{1}\left(\left|\frac{1}{2} - \xi_i\right| \leq \frac{1}{20}\right) \mathbb{1}\left(\left|\frac{1}{2} - \xi_j\right| \leq \frac{1}{20}\right)\right)\left(1 - \frac{1}{2}\left(\left|\frac{1}{2} - \xi_i\right| + \left|\frac{1}{2} - \xi_j\right|\right)\right)$.
The previous two functions had relatively low density (by construction, $\rho_n \leq 0.25$ for the dot product function and $\rho_n \leq 0.113$ for the horseshoe function). This final function has $\rho_n \leq 0.759$, allowing us to test the performance with higher density levels.

In the estimation procedure we use the normal kernel:[12] $K(u) = e^{-\frac{u^2}{2}}$ and the bandwidth $\hat{a}$ chosen by maximising $\ell(A, a_n)$, as described in Section 1.3.4. We look at sample sizes between $n = 100$ and $n = 1000$. We plot the heat maps showing the true function alongside the estimated ones.

Fig. 1.1 shows the comparison of the oracle value, the observed binary matrix, our method with the chosen bandwidth and with double the chosen bandwidth, our alternative method $\hat{h}^{(K2)}$, the estimator of Zhang, Levina, and Zhu (2017) and the dot product estimator used in Levin and Levina (2019) (with different dimensions k of the vector xi_i) at two different sample sizes.

The performance of all methods improves with sample size. Our methods at the default bandwidth tend to do a decent job, but they verge on the side of overfitting the observed

[12] The theoretical part of the paper imposes an assumption that the support of the kernel should be bounded. This is not satisfied for the normal kernel, but we chose not to rerun all the simulations to save computational cost. A smaller run of simulations comparing the performance with normal and quartic kernel $\left(K(u) = \frac{15}{16}\left(1 - u^2\right)^2 \mathbb{1}(|u| < 1)\right)$ confirmed that the choice of the kernel has minimal impact on the results. The choice of bandwidth is significantly more important.



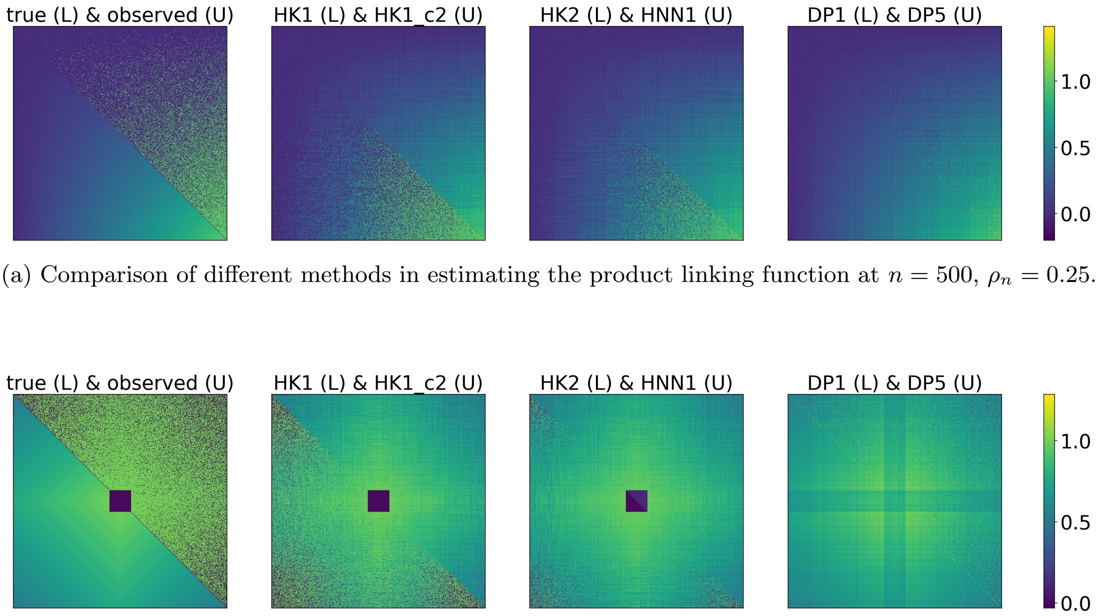
(a) Comparison of different methods in estimating the horseshoe function at $n = 100$, $\rho_n = 0.113$.

(b) Comparison of different methods in estimating the horseshoe function at $n = 500$, $\rho_n = 0.063$.

Figure 1.1: Comparison of different methods in estimating the horseshoe linking function. Each half of a square corresponds to a different matrix: 'true' is the oracle true values (first plot lower half), 'observed' is the binary matrix A used for estimation (first plot upper half), 'HK1' is our method based on the chosen bandwidth $\hat{a}$ (second plot lower half), 'HK1_c2' is our method based on double the chosen bandwidth $2\hat{a}$ (second plot upper half), 'HK2' is our alternative method based on the linking function estimator $\hat{h}^{(K2)}$ (third plot lower half), 'HNN1' uses the nearest neighbour estimator of Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size (third plot upper half), 'DPk' are the dot product estimators as in Levin and Levina (2019) based on assuming a k-dimensional ξ_i (fourth plot, lower half based on $k = 1$, upper half based on $k = 5$).

sample. For the larger sample size (Fig. 1.1b) the version which uses doubled bandwidth provides a better approximation than the default. The performance of $\hat{h}^{(K2)}$ is very similar to that of our main method. The estimator of Zhang, Levina, and Zhu (2017) is similar to ours with doubled bandwidth and it verges on the side of oversmoothing, which is causing some higher estimates for probabilities (the second arc above the main one). This gets better but is not completely fixed for the larger sample size. The dot product functions are not expected to perform well as they are misspecified, but the one with a larger dimension could provide a better approximation. We can see that at $k = 1$ the fit is very poor and at $k = 5$ it creates a pattern different from the true one. We have checked larger values (up to $k = 20$) and they tend to give similar results to $k = 5$. Note also that the dot product estimation is not bounded between 0 and 1, which explains why the estimates exceed these bounds.

Fig. 1.2 compares the same methods for other objective functions, both of which have denser connections. We can see that our method is again performing well when we use doubled bandwidth, but the original choice of $\hat{a}$ appears to be too low: the estimate is overfitting and too similar to the observed adjacency matrix. $\hat{h}^{(K2)}$ has a similar performance for the product



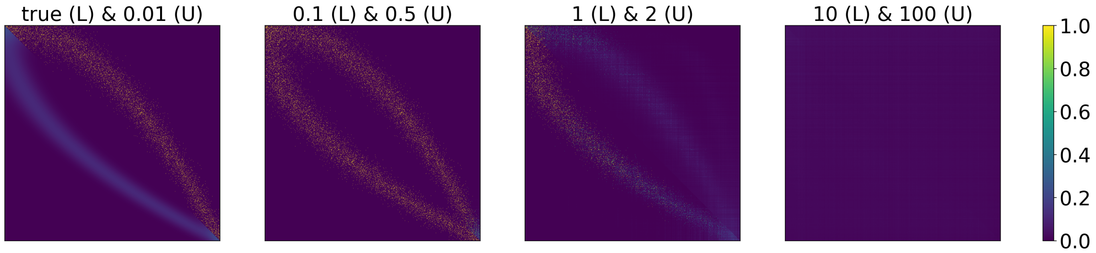
(a) Comparison of different methods in estimating the product linking function at $n = 500$, $\rho_n = 0.25$.

(b) Comparison of different methods in estimating the high density linking function at $n = 500$, $\rho_n = 0.759$.

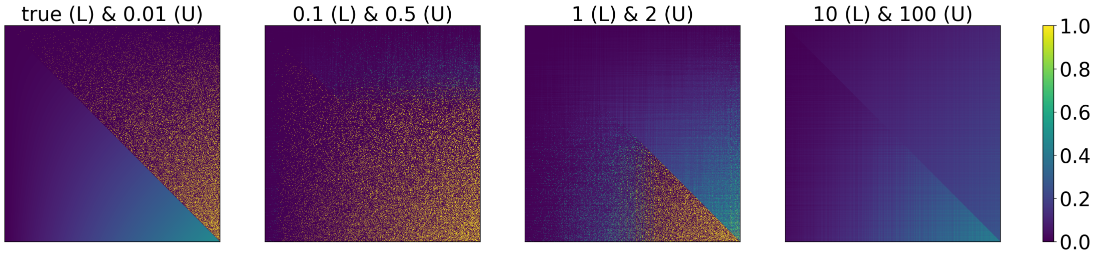
Figure 1.2: Comparison of different methods in estimating the linking functions. Each half of a square corresponds to a different matrix: 'true' is the oracle true values (first plot lower half), 'observed' is the binary matrix A used for estimation (first plot upper half), 'HK1' is our method based on the chosen bandwidth $\hat{a}$ (second plot lower half), 'HK1_c2' is our method based on double the chosen bandwidth $2\hat{a}$ (second plot upper half), 'HK2' is our alternative method based on the linking function estimator $\hat{h}^{(K2)}$ (third plot lower half), 'HNN1' uses the nearest neighbour estimator of Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size (third plot upper half), 'DPk' are the dot product estimators as in Levin and Levina (2019) based on assuming a k-dimensional ξ_i (fourth plot, lower half based on $k = 1$, upper half based on $k = 5$).

function and better for the high density linking function, avoiding the issue of overfitting at the chosen bandwidth. The estimator of Zhang, Levina, and Zhu (2017) performs very well in both cases. The dot product estimator is very good when correctly specified but does not have much advantage over ours with wider bandwidth or Zhang, Levina, and Zhu (2017)'s when we use incorrect $k = 5$. It does a surprisingly good job at some sections of the high density function estimation but it cannot handle the discontinuous jump. The higher $k = 5$ does not provide any advantage over $k = 1$.

Fig. 1.3 provides a comparison of our method at different bandwidths. The original choice is $c = 1$ and we can see that it makes a big difference in the performance of the procedure. At lower values the estimates tend to overfit the original observed adjacency matrix and lack the smoothness of the true function. Unfortunately, it appears that the choice based on our numerical optimisation procedure tends to be a bit too low: the wider bandwidths with $c = 2$ or even $c = 10$ (Fig. 1.3b) provide a better approximation to the true function. This is especially true when the fitted function is relatively simple: in those cases it is easier to find a close match to any observation, one which is close not only in terms of the underlying features but also



(a) Comparison of different bandwidths in estimating the horseshoe linking function at $n = 1000$, $\rho_n = 0.113$.



(b) Comparison of different bandwidths in estimating the product linking function at $n = 1000$, $\rho_n = 0.25$.

Figure 1.3: Comparison of different bandwidths in estimating the linking functions. Each half of a square corresponds to a different matrix: 'true' is the oracle true values (first plot lower half), the remaining plots look our estimator $\hat{h}$ using bandwidths of the form $c \times \hat{a}$ for $c \in \{0.01, 0.1, 0.5, 1, 2, 10, 100\}$ marked above the corresponding plots.

the realised randomness. This is why the method performs worse for the relatively simple dot product function while it does well at estimating the more challenging horseshoe function.

Possible solutions to this issue include using an alternative objective function for bandwidth selection which either adds a penalty term for lack of smoothness (provided that we expect the true linking function to be smooth) or use cross-validation to avoid overfitting. Both of these are promising but require choosing further parameters for the penalty size or the number of subsamples for cross-validation.

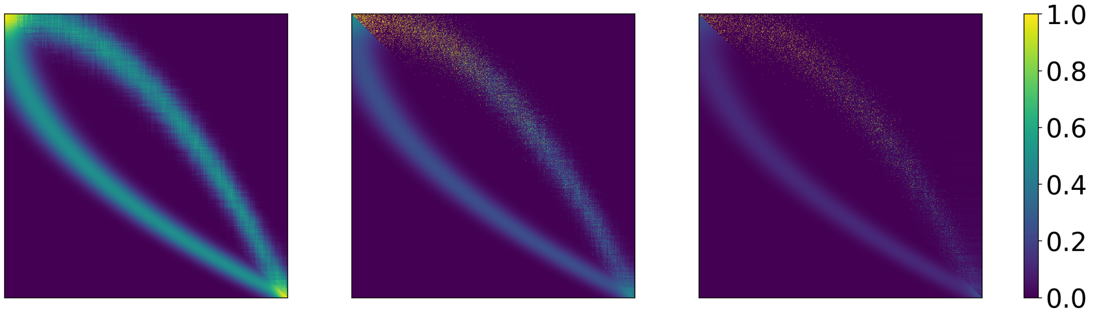


Figure 1.4: Comparison of different sparsity levels in estimating the horseshoe linking function at $n = 1000$. Each half of a square corresponds to a different matrix: the lower halves show the true oracle values, the upper halves show our estimator $\hat{h}$ using bandwidths $\hat{a}$. The first plot has $\rho_n = 0.113$, the second $\rho_n = 0.054$, the third $\rho_n = 0.026$.

Fig. 1.4 shows that the performance gets worse as the true model becomes more sparse, which is what we should expect. For the example with $n = 1000$ at $\rho_n = 0.113$ the performance is still very good, at lower values it gets worse but still recovers positive values in the correct locations, just with a tendency to overfit and predict more extreme values of probabilities (close to 0 or 1, depending on the realised value). This likely happens because when there are few observations that provide a good match the procedure relies on giving a high weight to the original observation only, without relying on close neighbours.

1.6 Conclusion and Extensions

In this paper we propose a nonparametric linking probabilities estimator and provide conditions for its uniform consistency.

This paper is a contribution in its own right, but it is also intended to provide a building block for further projects. Having a general framework for estimating the key component capturing the dependence structure in a network can be useful for replicating a similar dependence structure, which can be used for bootstrapping network data. In models where we believe the latent linking probability is a determinant of agents' behaviour, we may be able to treat the estimated linking probabilities as a proxy to form models with lower bias than those using only binary information on the observed links.

An important extension would be to consider estimation with observed covariates, like in the case of Kitamura and Laage (2024). Natural ways of implementing this would involve firstly restricting attention to individuals with similar observable characteristics to those whose probability of match we wish to estimate, and then applying our procedure on the subgraph, or using our procedure with a distance which is a weighted average of our current distance and one based on similarity in observables. Apart from complicating the technical derivation of the results, this modification could require higher computational costs and observing higher sample sizes to ensure the same quality of estimation (it would be harder to find good matches in terms of both observables and network position). We are not aware of results similar to the Aldous-Hoover representation theorem for the case with observable covariates (perhaps conditional on covariates), which could limit the applicability of the method. Developing these kinds of representations or showing why they do not arise would be another valuable extension.

Another possible direction would be exploring extensions of this model to the directed case: for this we would not average the two one-sided estimates. These one-sided estimates could also be informative about signals sent by one agent to another that determine links in models of strategic network formation.

Bibliography for Chapter 1

Abadie, Alberto, and Guido W Imbens. 2008. "On the failure of the bootstrap for matching estimators." *Econometrica* 76 (6): 1537–1557.

Aldous, David J. 1981. "Representations for partially exchangeable arrays of random variables." *Journal of Multivariate Analysis* 11 (4): 581–598.

Amini, Arash A, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. 2013. "Pseudo-likelihood methods for community detection in large sparse networks."

Athreya, Avanti, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. 2018. "Statistical inference on random dot product graphs: a survey." *Journal of Machine Learning Research* 18 (226): 1–92.

Auerbach, Eric. 2022. "Identification and Estimation of a Partially Linear Regression Model Using Network Data." *Econometrica* 90 (1): 347–365.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2013. "The diffusion of microfinance." *Science* 341 (6144): 1236498.

Breza, Emily, and Arun G Chandrasekhar. 2019. "Social networks, reputation, and commitment: evidence from a savings monitors experiment." *Econometrica* 87 (1): 175–216.

Carter, Michael, Rachid Laajaj, and Dean Yang. 2021. "Subsidies and the African Green Revolution: direct effects and social network spillovers of randomized input subsidies in Mozambique." *American Economic Journal: Applied Economics* 13 (2): 206–229.

Chetty, Raj, Matthew O Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, et al. 2022. "Social capital I: measurement and associations with economic mobility." *Nature* 608 (7921): 108–121.

Green, Alden, and Cosma Rohilla Shalizi. 2022. "Bootstrapping exchangeable random graphs." *Electronic Journal of Statistics* 16 (1): 1058–1095.

Guédon, Olivier, and Roman Vershynin. 2016. "Community detection in sparse networks via Grothendieck's inequality." *Probability Theory and Related Fields* 165 (3): 1025–1049.

Hoover, Douglas N. 1979. "Relations on Probability Spaces and Arrays of." *t, Institute for Advanced Study.*

Kallenberg, Olav. 1989. "On the representation theorem for exchangeable arrays." *Journal of Multivariate Analysis* 30 (1): 137–154.

Kitamura, Yuichi, and Louise Laage. 2024. *Estimating Stochastic Block Models in the Presence of Covariates.*

Lei, Jing, and Alessandro Rinaldo. 2015. "Consistency of spectral clustering in stochastic block models." *The Annals of Statistics,* 215–237.

Levin, Keith, and Elizaveta Levina. 2019. "Bootstrapping networks with latent space structure." *arXiv preprint arXiv:1907.10821.*

Lovász, László. 2012. *Large networks and graph limits.* Vol. 60. American Mathematical Soc.

Orbanz, Peter, and Daniel M Roy. 2014. "Bayesian models of graphs, arrays and other exchangeable random structures." *IEEE transactions on pattern analysis and machine intelligence* 37 (2): 437–461.

Rohe, Karl, Sourav Chatterjee, and Bin Yu. 2011. "Spectral clustering and the high-dimensional stochastic block model."

Wang, Lawrence. 2016. "Network Comparisons using Sample Splitting." *PhD thesis, Carnegie Mellon University.*

Zeleneev, Andrei. 2020. "Identification and estimation of network models with nonparametric unobserved heterogeneity." *Department of Economics, Princeton University.*

Zhang, Yuan, Elizaveta Levina, and Ji Zhu. 2017. "Estimating network edge probabilities by neighbourhood smoothing." *Biometrika* 104 (4): 771–783.

Appendix

List of all notation

The notation in this file can get a bit heavy so we provide this list for reference.

- n – sample size, number of individuals in the network.
- A – an $n \times n$ adjacency matrix. Binary, symmetric, observed.
- A_{ij} – i, jth entry of the matrix A: 1 if i, j are connected (are neighbours), 0 if they are not.
- i, j, k, s, t – usually used to refer to one of the n individuals.
- ξ_i – vector of characteristics of individual i, enters the linking function.
- F_0 – distribution of ξ_i.
- $h_{0,n}$ – linking function, takes characteristics ξ_i, ξ_j as inputs and outputs the probability with which individuals i and j are linked. If the inputs are vectors $\xi(\iota) = (\xi_{\iota 1}, \xi_{\iota 2}, \dots, \xi_{\iota m})$ of characteristics of multiple individuals it outputs the matrix of linking probabilities.
- ρ_n – density/sparsity parameter. Density in the sense that it is the expected edge density, sparsity in the sense that as $n \to \infty$ the density of edges decreases: $\rho_n \to 0$.
- w_0 – underlying linking probability before accounting for sparsity: $\rho_n w_0 = h_{0,n}$.
- $\varphi(\xi_i, \xi_t) = E\left(\frac{A_{is}A_{ts}}{\rho_n^2} | \xi_i, \xi_t\right)$ – a function measuring the probability of a common friend between i and j normalised by the sparsity level.
- $d_{ij} = \sqrt{E\left(E\left(w_0\left(\xi_t, \xi_s\right)\left(w_0\left(\xi_i, \xi_s\right) - w_0\left(\xi_j, \xi_s\right)\right) \middle| \, \xi_i, \xi_j, \xi_t\right)^2 \middle| \, \xi_i, \xi_j \right)}$ – theoretical distance between i and j.
- $\hat{d}_{ij} = \frac{1}{\rho_n^2} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\frac{1}{n} \sum_{s=1}^{n} A_{ts}\left(A_{is} - A_{js}\right)\right)^2}$ – estimated distance between i and j.

- $\hat{h}_n$ – estimated linking function:

$$\hat{h}_n(\xi_i, \xi_j) = \frac{\tilde{h}_n(\xi_i, \xi_j) + \tilde{h}_n(\xi_j, \xi_i)}{2} \quad \text{where} \quad \tilde{h}_n(\xi_i, \xi_j) = \frac{\sum_{\substack{t=1 \\ t\neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) A_{tj}}{\sum_{\substack{t=1 \\ t\neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right)}$$

- K – kernel function used in estimating linking probability.

- a_n – a bandwidth parameter, chosen by the researcher.

- $\hat{}$ – an estimate.

- $\max_{i,j} \equiv \max_{i,j\in\{1,2,\ldots,n\}}$ – maximum over indices in a specific sample of size n.

- $\max_{\xi_i} \equiv \max_{\xi_i \in Supp(\xi_i)}$ – maximum over all $\xi_i \in Supp(\xi_i)$.

- $N(\xi_j, \delta) = \left\{\xi_k : \sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta\right\}$ – the neighbourhood of ξ_j of size δ.

- $\omega(\delta) = \inf_{\xi_j \in Supp(\xi_j)} P\left(\xi_k \in N(\xi_j, \delta) | \, \xi_j\right)$ – the infimum over all possible ξ_i of the measures of their neighbourhoods of size δ.

- $b_n = \frac{a_n}{\rho_n^4}$ – a bandwidth parameter normalised by sparsity; the effective bandwidth size after accounting for the rate at which density goes to zero.

- $\hat{h}_n^-$ – leave-one-out version of $\hat{h}_n$, evaluated in the same way as $\hat{h}_n$ but without the observations $t = i, j$. Used for numerically choosing the optimal bandwidth.

- $\ell(A, a_n)$ – log-likelihood used for numerically choosing the optimal bandwidth. Defined in Eq. (1.12).

- $\hat{a}$ – numerically chosen optimal bandwidth. Defined in Eq. (1.13).

- η – a vector of random variables which together with the linking function determine the realised links in A. We assume $\eta_{ij} \overset{ind}{\sim} \mathcal{U}[0, 1]$ for $1 \leq i \leq j \leq n$ and η independent of ξ.

- C – generic positive constant, its value may change between different expressions in which it is used.

- C_ε – a positive constant which depends on $\varepsilon > 0$. Its value may change between different expressions in which it is used.

- T_n – a remainder term used in the proof of Theorem 1.

- M_w – an upper bound on the value of w_0: $\sup_{\xi_i, \xi_j} |w_0(\xi_i, \xi_j)| \leq M_w$.

- $r_n(i) = E\left(K\left(\frac{d_{it}^2}{b_n}\right) \middle| \, \xi_i\right)$ – the shorthand notation for the expected kernel weights based on the distance between i and other individuals used in the estimation of $\hat{h}_n\left(\xi_i, \xi_j\right)$.

- $\hat{r}_n(i) = \frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n} K\left(\frac{\hat{d_{it}^2}}{b_n}\right)$ – the estimate of $r_n(i)$.
- $r_n = \inf_{\xi_i} r_n(i)$ – the smallest possible expected kernel weight. We need to ensure it is not too small or we would not be able to successfully estimate $h_{0,n}\left(\xi_i, \xi_j\right)$.

Appendix 1.A Proofs

Subsection 1.A.1 Proof of uniform consistency of the linking function estimator

Proof of Theorem 1. Throughout this argument we use C_ε to denote a positive constant which depends on $\varepsilon > 0$. The value of C_ε may change between different expressions in which it is used.

By definition,

$$\max_{i,j}\left|\frac{\hat{h}_n(\xi_i,\xi_j) - h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right| = \max_{i,j}\left|\frac{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^n K\left(\frac{\hat{d}_{it}^2}{b_n}\right)\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^n K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|$$

$$= \max_{i,j}\left|\frac{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^n K\left(\frac{\hat{d}_{it}^2}{b_n}\right)\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)}{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right) + \left(\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^n K\left(\frac{\hat{d}_{it}^2}{b_n}\right) - E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)\right)}\right|$$

$$\leq \left(\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^n \frac{K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)\right|\right)\left(1+\max_{i,j}\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^n K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|\right) \tag{1.14}$$

The inequality follows from $\max\left|\frac{a}{b+c}\right| \leq \max\left|\frac{a}{b}\right| + \max\left|\frac{ac}{b(b+c)}\right| \leq \left(\max\left|\frac{a}{b}\right|\right)\left(1+\max\left|\frac{c}{b+c}\right|\right)$, where $b = E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)$. In Lemma 1.A.1 we show that the second factor converges almost surely to one.

We now focus on the first factor. Since $K(\cdot)$ is Lipschitz continuous with a Lipschitz constant C (by Assumption 1.3):

$$K\left(\frac{\hat{d}_{it}^2}{b_n}\right) \leq K\left(\frac{d_{it}^2}{b_n}\right) + \left|K\left(\frac{\hat{d}_{it}^2}{b_n}\right) - K\left(\frac{d_{it}^2}{b_n}\right)\right| \leq K\left(\frac{d_{it}^2}{b_n}\right) + C\left|\frac{\hat{d}_{it}^2 - d_{it}^2}{b_n}\right|.$$

It follows that

$$\begin{aligned}
&\max_{i,j}\left|\frac{1}{(n-1)r_n(i)}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)\right|\\
&\le \max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)+C\left|\frac{\hat{d}_{it}^2-d_{it}^2}{b_n r_n(i)}\right|\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)\right|\\
&\le \max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)\right|\\
&\quad + C\max_{i,j,t,t\neq j}\left|\frac{\hat{d}_{it}^2-d_{it}^2}{b_n r_n(i)}\right|\underbrace{\max_{i,j}\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\left|\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right|}_{\le 2M_w+\underbrace{T_n}_{\xrightarrow{a.s.}0}}.
\end{aligned}$$

$M_w<\infty$ is defined in Lemma 1.A.1. In Lemma 1.A.2 we define T_n and show that the last factor in the last expression is almost surely bounded, hence

$$\begin{aligned}
&\max_{i,j}\left|\frac{1}{(n-1)\rho_n r_n(i)}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)(A_{tj}-h_{0,n}(\xi_i,\xi_j))\right|\\
&\quad\le \max_{i,j}\left|\frac{1}{(n-1)\rho_n r_n(i)}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{d_{it}^2}{b_n}\right)(A_{tj}-h_{0,n}(\xi_t,\xi_j)+h_{0,n}(\xi_t,\xi_j)-h_{0,n}(\xi_i,\xi_j))\right|\\
&\qquad + C(2M_w+T_n)\frac{1}{b_n r_n}\left(\max_{i,j}\left|\hat{d}_{ij}^2-d_{ij}^2\right|\right)\\
&\quad\le \max_{i,j}\left|\frac{\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{d_{it}^2}{b_n}\right)(A_{tj}-h_{0,n}(\xi_t,\xi_j))}{(n-1)\rho_n r_n(i)}\right|\\
&\qquad + \max_{i,j}\left|\frac{\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{d_{it}^2}{b_n}\right)(h_{0,n}(\xi_t,\xi_j)-h_{0,n}(\xi_i,\xi_j))}{(n-1)\rho_n r_n(i)}\right|\\
&\qquad + C(2M_w+T_n)\frac{1}{b_n r_n}\left(\max_{i,j}\left|\hat{d}_{ij}^2-d_{ij}^2\right|\right).
\end{aligned}$$

where $T_n \xrightarrow{a.s.} 0$. We complete the proof by showing that the three terms go to zero in probability in Lemma 1.A.3, Lemma 1.A.4 and Lemma 1.A.5. □

Lemma 1.A.1. *Under the assumptions of Theorem 1, for any* $\varepsilon > 0$*:*

$$\sum_{n=3}^{\infty} P\left(\left|\max_{i,j}\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|\right|>\varepsilon\right)$$
$$=O\left(\sum_{n=3}^{\infty}n^2\exp\left(-nr_nC_\varepsilon\right)+\sum_{n=3}^{\infty}n^4\exp\left(-nb_n^2r_n^2\rho_n^2C_\varepsilon\right)\right)=O(1).$$

hence

$$\max_{i,j}\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|\xrightarrow{a.s.}0.$$

Proof. Take any $\varepsilon > 0$. We start by using a union bound:

$$P\left(\max_{i,j}\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|>\varepsilon\right)\leq n^2P\left(\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)|\xi_i,\xi_j\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|>\varepsilon\right)$$

Let $r_n(i)=E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)\geq r_n\geq 0$ and $\hat{r}_n(i)=\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)$. We have:

$$\begin{aligned}P\left(\left|1-\frac{r_n(i)}{\hat{r}_n(i)}\right|>\varepsilon\right)&\leq P\left(|\hat{r}_n(i)-r_n(i)|>\varepsilon|\hat{r}_n(i)|\right)\\&\leq P\left(|\hat{r}_n(i)-r_n(i)|>\varepsilon|\hat{r}_n(i)|\text{ and }|\hat{r}_n(i)|\geq\frac{r_n(i)}{2}\right)\\&\quad+P\left(|\hat{r}_n(i)-r_n(i)|>\varepsilon|\hat{r}_n(i)|\text{ and }|\hat{r}_n(i)|<\frac{r_n(i)}{2}\right)\\&\leq P\left(|\hat{r}_n(i)-r_n(i)|>\varepsilon\frac{r_n(i)}{2}\right)+P\left(|\hat{r}_n(i)|<\frac{r_n(i)}{2}\right)\\&\leq P\left(\left|\frac{\hat{r}_n(i)}{r_n(i)}-1\right|>\frac{\varepsilon}{2}\right)+P\left(\left|\frac{\hat{r}_n(i)}{r_n(i)}-1\right|>\frac{1}{2}\right)\end{aligned}$$

where the last line follows from:

$$\begin{aligned}P\left(|\hat{r}_n(i)|<\frac{r_n(i)}{2}\right)&\leq P\left(\hat{r}_n(i)<\frac{r_n(i)}{2}\right)=P\left(\hat{r}_n(i)-r_n(i)<-\frac{r_n(i)}{2}\right)\\&=P\left(r_n(i)-\hat{r}_n(i)>\frac{r_n(i)}{2}\right)\leq P\left(|r_n(i)-\hat{r}_n(i)|>\frac{r_n(i)}{2}\right)\\&\leq P\left(\left|\frac{\hat{r}_n(i)}{r_n(i)}-1\right|>\frac{1}{2}\right).\end{aligned}$$

We use the above derivation and the law of iterated expectations to get am upper bound of the form:

$$P\left(\max_{i,j}\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\Big|\,\xi_i\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right|>\varepsilon\right)\leq n^2E\left(P\left(\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}{r_n(i)}-1\right|>\frac{\varepsilon}{2}\middle|\,\xi_i,\xi_j\right)\right)$$
$$+\,n^2E\left(P\left(\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}{r_n(i)}-1\right|>\frac{1}{2}\middle|\,\xi_i,\xi_j\right)\right).$$

The last two terms are identical, up to the value of ε. We use $K(\cdot)$ Lipschitz continuous and separate out the terms with $t=i,j$, so that the remaining average is of i.i.d terms that only depend on t.

$$n^2E\left(P\left(\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}{r_n(i)}-1\right|>\frac{\varepsilon}{2}\middle|\,\xi_i,\xi_j\right)\right)$$
$$\leq n^2E\left(P\left(\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}-1+\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}C\left|\frac{\hat{d}_{it}^2-d_{it}^2}{r_n(i)b_n}\right|\right|>\frac{\varepsilon}{2}\middle|\,\xi_i,\xi_j\right)\right)$$
$$\leq n^2E\left(P\left(\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}-1\right|>\frac{\varepsilon}{4}\middle|\,\xi_i,\xi_j\right)\right)$$
$$+\,n^2E\left(P\left(\left|\max_{i,t}\left(\frac{\hat{d}_{it}^2-d_{it}^2}{r_n(i)b_n}\right)\right|>\frac{\varepsilon}{4C}\middle|\,\xi_i,\xi_j\right)\right)$$
$$\leq n^2E\left(P\left(\left|\frac{1}{n-2}\sum_{\substack{t=1\\t\neq i,j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}-1\right|>\frac{\varepsilon}{4}-\frac{2C}{(n-2)r_n}\middle|\,\xi_i,\xi_j\right)\right)+O\left(n^4\exp\left(-nb_n^2r_n^2\rho_n^2C_\varepsilon\right)\right)$$

where the last rate follows from Lemma 1.A.5. For the first term we apply Bernstein's inequality: conditional on ξ_i,ξ_j, $\frac{1}{r_n(i)}K\left(\frac{d_{it}^2}{b_n}\right)-1$ are i.i.d., mean zero, bounded by $\frac{2C}{r_n}$, with variance $O\left(\frac{1}{r_n}\right)$:

$$Var\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}-1\middle|\,\xi_i\right)\leq\frac{E\left(\left(K\left(\frac{d_{it}^2}{b_n}\right)\right)^2\Big|\,\xi_i\right)}{r_n(i)^2}\leq\frac{CE\left(K\left(\frac{d_{it}^2}{b_n}\right)\Big|\,\xi_i\right)}{\left(E\left(K\left(\frac{d_{it}^2}{b_n}\right)\Big|\,\xi_i\right)\right)^2}$$
$$=\frac{C}{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\Big|\,\xi_i\right)}\leq\frac{C}{r_n}=O\left(\frac{1}{r_n}\right),$$

hence for any $\varepsilon > 0$:

$$
\begin{aligned}
& n^2 P\left(\left|\frac{1}{n-2}\sum_{\substack{t=1\\ t\neq i,j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}-1\right| > \frac{\varepsilon}{4}-\frac{2C}{(n-2)r_n}\right|\xi_i,\xi_j\right) \\
& \leq 2n^2 \exp\left(-\frac{(n-2)\left(\frac{\varepsilon}{4}-\frac{2C}{(n-2)r_n}\right)^2}{2\left(O\left(\frac{1}{r_n}\right)+\frac{1}{3}\frac{C}{r_n}\left(\frac{\varepsilon}{4}-\frac{2C}{(n-2)r_n}\right)\right)}\right) \\
& \leq n^2 \exp\left(-nr_nC_\varepsilon\right).
\end{aligned}
$$

where $C_\varepsilon > 0$ is some constant dependent on ε. Note that the final value does not depend on the choice of ξ_i, ξ_j, hence it does not change when we take expectation over ξ_i, ξ_j.

It remains to show that the following expression:

$$
\begin{aligned}
& \sum_{n=3}^{\infty} P\left(\max_{i,j}\left|1-\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{\frac{1}{n-1}\sum_{\substack{t=1\\ t\neq j}}^{n}K\left(\frac{\hat{d}_{it}^2}{b_n}\right)}\right| > \varepsilon\right) \\
& \leq O\left(\sum_{n=3}^{\infty} n^2\exp\left(-nr_nC_\varepsilon\right)+\sum_{n=3}^{\infty} n^4\exp\left(-nb_n^2r_n^2\rho_n^2C_\varepsilon\right)\right)
\end{aligned}
$$

is bounded. The last sum is bounded by arguments shown in Lemma 1.A.5. For the first sum we have:

$$
\sum_{n=3}^{\infty} n^4 e^{-nr_nC_\varepsilon} = \sum_{n=3}^{\infty} n^2 e^{-nr_nC_\varepsilon\log(n)\frac{1}{\log(n)}} = \sum_{n=3}^{\infty} n^2\left(e^{\log(n)}\right)^{-C_\varepsilon\frac{nr_n}{\log(n)}} = \sum_{n=3}^{\infty} n^{2-C_\varepsilon\frac{nr_n}{\log(n)}}.
$$

It remains to show $\frac{nr_n}{\log(n)} \to \infty$. We start by showing $r_n \geq Cb_n^{\frac{1}{2\alpha}}$:

$$
\begin{aligned}
r_n = \inf_{\xi_i} r_n(i) = \inf_{\xi_i} E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right) &\geq C_1 \inf_{\xi_i} P\left(\frac{d_{it}^2}{b_n}\leq C_2\middle|\xi_i\right) \\
&\geq C_1 \inf_{\xi_i} P\left(E_s\left(\left(w_0\left(\xi_i,\xi_s\right)-w_0\left(\xi_t,\xi_s\right)\right)^2\middle|\xi_i,\xi_t\right)\leq\frac{C_2}{M_w^2}b_n\middle|\xi_i\right) \\
&\geq C_1 \inf_{\xi_i} P\left(\xi_t\in N\left(\xi_i,\sqrt{\frac{C_2}{M_w^2}b_n}\right)\middle|\xi_i\right) = C_1\omega\left(\sqrt{\frac{C_2}{M_w^2}b_n}\right) \geq Cb_n^{\frac{1}{2\alpha}}.
\end{aligned}
$$

In the first inequality we use the part of Assumption 1.3 which says the kernel is separated from 0 for input values sufficiently close to 0. The second inequality comes from:

$$
\begin{aligned}
d_{ij}^2 &= E_t\left(\left(E_s\left(w_0\left(\xi_t,\xi_s\right)\left(w_0\left(\xi_i,\xi_s\right)-w_0\left(\xi_j,\xi_s\right)\right)\middle|\xi_i,\xi_j,\xi_t\right)\right)^2\middle|\xi_i,\xi_j\right) \\
&\leq E_t\left(E_s\left(w^2\left(\xi_t,\xi_s\right)\left(w_0\left(\xi_i,\xi_s\right)-w_0\left(\xi_j,\xi_s\right)\right)^2\middle|\xi_i,\xi_j,\xi_t\right)\middle|\xi_i,\xi_j\right) \\
&\leq M_w^2 E_s\left(\left(w_0\left(\xi_i,\xi_s\right)-w_0\left(\xi_j,\xi_s\right)\right)^2\middle|\xi_i,\xi_j\right) \leq M_w^4 < \infty
\end{aligned}
$$

where the first inequality is due to Jensen's inequality and the second follows from the fact that for any $\xi_i, \xi_j \in Supp(\xi)$ $w_0(\xi_i, \xi_j)$ is bounded; we denote the bound by $M_w < \infty$. To see this, recall that $\rho_n w_0(u, v) = h_{0,n}(u, v) \in [0, 1]$, hence we have $w_0(u, v) \in \left[0, \frac{1}{\rho_n}\right]$ for all $n \in \mathbb{N}$. Then also $w_0(u, v) \in \bigcap_{n=1}^{\infty} \left[0, \frac{1}{\rho_n}\right] \subset \left[0, \frac{1}{\sup_n \rho_n}\right]$. $\sup_n \rho_n$ exists since ρ_n, which can be interpreted as the marginal probability of an edge, is bounded above by 1. Let $M_w = \frac{1}{\sup_n \rho_n}$ denote the upper bound on the size of w_0, i.e. for any $\xi_i, \xi_j \in Supp(\xi)$ we have $|w_0(\xi_i, \xi_j)| \leq M_w$.

The third inequality follows from the fact that if for some ξ_t we have $\sup_{\xi_s} |w_0(\xi_t, \xi_s) - w_0(\xi_i, \xi_s)| < \delta$, then $E_s \left((w_0(\xi_i, \xi_s) - w_0(\xi_t, \xi_s))^2 \middle| \xi_i, \xi_t \right) < \delta^2$, i.e. $\xi_t \in N(\xi_i, \delta)$. For the final steps we use Assumption 1.2. The required divergence follows from Assumption 1.4:

$$\frac{nr_n}{\log(n)} \geq C \frac{nb_n^{\frac{1}{2\alpha}}}{\log(n)} \to \infty.$$

□

Lemma 1.A.2. *Under the assumptions of Theorem 1, there exists a sequence of random variables T_n such that*

$$\sum_{n=3}^{\infty} P(|T_n| > \varepsilon) = O\left(\sum_{n=3}^{\infty} n^2 \exp(-n\rho_n C_\varepsilon) \right) = O(1) \quad \text{hence} \quad T_n \xrightarrow{a.s.} 0$$

and

$$\max_{i,j} \frac{1}{n-1} \sum_{\substack{t=1 \\ t \neq j}}^{n} \left| \frac{A_{tj} - h_{0,n}(\xi_i, \xi_j)}{\rho_n} \right| \leq 2M_w + T_n.$$

Proof. We start by looking at a representative term inside the summation.

$$\begin{aligned} \left| \frac{A_{tj} - h_{0,n}(\xi_i, \xi_j)}{\rho_n} \right| &\leq \left| \frac{A_{tj}}{\rho_n} \right| + \left| \frac{h_{0,n}(\xi_i, \xi_j)}{\rho_n} \right| \\ &= \frac{A_{tj}}{\rho_n} + w_0(\xi_i, \xi_j) \\ &= \frac{A_{tj}}{\rho_n} - w_0(\xi_t, \xi_j) + w_0(\xi_t, \xi_j) + w_0(\xi_i, \xi_j) \\ &\leq \frac{A_{tj}}{\rho_n} - w_0(\xi_t, \xi_j) + 2M_w. \end{aligned}$$

We use triangle inequality, the fact that A_{tj} and $h_{0,n}(\xi_i, \xi_j)$ are non-negative and the definition of $w_0(\xi_i, \xi_j)$. We add and subtract $w_0(\xi_t, \xi_j)$ and use the fact that all possible values of w_0 are bounded by M_w.

Going back to the sum:

$$\max_{i,j} \frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n} \left|\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right| \leq 2M_w + \max_{i,j} \frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{A_{tj}}{\rho_n} - w_0(\xi_t,\xi_j)$$

$$\leq 2M_w + \max_{i,j,i\neq j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{A_{tj}}{\rho_n} - w_0(\xi_t,\xi_j)\right| + \max_i \left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq i}}^{n}\frac{A_{ti}}{\rho_n} - w_0(\xi_t,\xi_i)\right| = 2M_w + T_n.$$

In the second step we split into cases with $i \neq j$ and $i = j$. We apply union bound and Bernstein's theorem to the averages. For the first one, we separate out the term with $t = i$ (we later condition on ξ_i, ξ_j, we want the remaining terms in the sum to be i.i.d. after conditioning). $\frac{A_{tj}}{\rho_n} - w_0(\xi_t,\xi_j)$ for $t \neq i, j$ are, conditional on ξ_i, ξ_j, independent, zero mean:

$$E\left(\frac{A_{tj}}{\rho_n} - w_0(\xi_t,\xi_j)\middle|\,\xi_i,\xi_j\right) = E\left(E\left(\frac{A_{tj}}{\rho_n} - w_0(\xi_t,\xi_j)\middle|\,\xi_i,\xi_j,\xi_t\right)\middle|\,\xi_i,\xi_j\right)$$
$$= E\left(E\left(\frac{A_{tj}}{\rho_n}\middle|\,\xi_i,\xi_j,\xi_t\right) - w_0(\xi_t,\xi_j)\middle|\,\xi_i,\xi_j\right) = E\left(w_0(\xi_t,\xi_j) - w_0(\xi_t,\xi_j)|\,\xi_i,\xi_j\right) = 0$$

and bounded by $\frac{1}{\rho_n}$: since A and $h_{0,n}$ take values in $[0,1]$, we have
$\left|\frac{A_{tj}}{\rho_n} - w_0(\xi_t,\xi_j)\right| = \left|\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right| \leq \frac{1}{\rho_n}$. The second moments are $O\left(\frac{1}{\rho_n}\right)$:

$$Var\left(\frac{1}{\rho_n}\left(A_{tj} - h_{0,n}(\xi_t,\xi_j)\right)\middle|\,\xi_i,\xi_j\right) = E\left(\left(\frac{1}{\rho_n}\left(A_{tj} - h_{0,n}(\xi_t,\xi_j)\right)\right)^2\middle|\,\xi_i,\xi_j\right)$$
$$= E\left(\frac{1}{\rho_n^2}\left(h_{0,n}(\xi_t,\xi_j)\left(1 - h_{0,n}(\xi_t,\xi_j)\right)\right)\middle|\,\xi_i,\xi_j\right) = E\left(w_0(\xi_t,\xi_j)\left(\frac{1}{\rho_n} - w_0(\xi_t,\xi_j)\right)\middle|\,\xi_j\right)$$
$$= O\left(\frac{1}{\rho_n}\right) + O(1) = O\left(\frac{1}{\rho_n}\right).$$

For any $\varepsilon > 0$:

$$
\begin{aligned}
&P\left(\max_{i,j,i\neq j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{A_{tj}}{\rho_n}-w_0(\xi_t,\xi_j)\right|>\varepsilon\right)\\
&\leq P\left(\max_{i,j,i\neq j}\left|\frac{1}{n-2}\sum_{\substack{t=1\\t\neq i,j}}^{n}\frac{A_{tj}}{\rho_n}-w_0(\xi_t,\xi_j)\right|>\varepsilon-\frac{1}{(n-2)\rho_n}\right)\\
&\leq n(n-1)E\left(P\left(\left|\frac{1}{n-2}\sum_{\substack{t=1\\t\neq i,j}}^{n}\frac{A_{tj}}{\rho_n}-w_0(\xi_t,\xi_j)\right|>\varepsilon-\frac{1}{(n-2)\rho_n}\middle|\xi_i,\xi_j\right)\right)\\
&\leq 2n(n-1)\exp\left(-\frac{(n-2)\left(\varepsilon-\frac{1}{(n-2)\rho_n}\right)^2}{2\left(O\left(\frac{1}{\rho_n}\right)+\frac{1}{3}\frac{1}{\rho_n}\left(\varepsilon-\frac{1}{(n-2)\rho_n}\right)\right)}\right)\\
&\leq n^2\exp\left(-n\rho_n C_\varepsilon\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&P\left(\max_{i}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq i}}^{n}\frac{A_{ti}}{\rho_n}-w_0(\xi_t,\xi_i)\right|>\varepsilon\right)\\
&\leq nE\left(P\left(\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq i}}^{n}\frac{A_{ti}}{\rho_n}-w_0(\xi_t,\xi_i)\right|>\varepsilon\middle|\xi_i\right)\right)\\
&\leq 2n\exp\left(-\frac{(n-1)\varepsilon^2}{2\left(O\left(\frac{1}{\rho_n}\right)+\frac{1}{3}\frac{1}{\rho_n}\varepsilon\right)}\right)\\
&\leq n\exp\left(-n\rho_n C_\varepsilon\right).
\end{aligned}
$$

This is dominated by the previous term. Combining the above results, for any $\varepsilon > 0$:

$$
\begin{aligned}
\sum_{n=3}^{\infty}P\left(\max_{i,j}\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\left|\frac{A_{tj}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right|>2M_w+\varepsilon\right)&\leq\sum_{n=3}^{\infty}P\left(|T_n|>\varepsilon\right)\\
&\leq O\left(\sum_{n=3}^{\infty}n^2\exp\left(-n\rho_n C_\varepsilon\right)\right)<\infty.
\end{aligned}
$$

For the last claim, note that under Assumption 1.1 we have $\frac{\log(n)}{\rho_n n}\to 0$. Then:

$$
\sum_{n=3}^{\infty}n^2e^{-n\rho_n C_\varepsilon}=\sum_{n=3}^{\infty}n^2e^{-n\rho_n C_\varepsilon\log(n)\frac{1}{log(n)}}=\sum_{n=3}^{\infty}n^2\left(e^{\log(n)}\right)^{-C_\varepsilon\frac{\rho_n n}{\log(n)}}=\sum_{n=3}^{\infty}n^{2-C_\varepsilon\frac{\rho_n n}{\log(n)}}<\infty
$$

for any $C_\varepsilon > 0$, since $2 - C_\varepsilon \frac{\rho_n n}{\log(n)} \to -\infty$. Hence $T_n \xrightarrow{a.s} 0$ and the term of interest is almost surely bounded above by $2M_w$.

□

Lemma 1.A.3. *Under the assumptions of Theorem 1, for any $\varepsilon > 0$:*

$$\sum_{n=3}^{\infty} P\left(\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj} - h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right| > \varepsilon\right)$$
$$\leq O\left(\sum_{n=3}^{\infty} n^2 \exp\left(-nr_n\rho_n C_\varepsilon\right)\right) < \infty$$

hence

$$\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj} - h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right| \xrightarrow{a.s.} 0.$$

Proof. We start by separating the cases when $i \neq j$ and $i = j$:

$$\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj} - h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right|$$
$$\leq \max_{i,j,i\neq j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj} - h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right|$$
$$+ \max_{i}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq i}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{ti} - h_{0,n}(\xi_t,\xi_i)}{\rho_n}\right)\right|.$$

We split the first sum into the term with $t = i$ and the rest the rest (which is i.i.d. over t), use the triangle inequality and the fact that A_{ij} and $h_{0,n}(\xi_i,\xi_j)$ take values in $[0,1]$ for any choice of i, j while the kernel function is absolutely bounded.

$$\max_{i,j,i\neq j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right|$$

$$\leq \max_{i,j,i\neq j}\left|\frac{n-2}{n-1}\frac{1}{n-2}\sum_{\substack{t=1\\t\neq i,j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right|$$

$$+\max_{i,j,i\neq j}\frac{1}{n-1}\left(\underbrace{\left|\frac{K\left(\frac{d_{ii}^2}{b_n}\right)}{r_n(i)}\right|}_{\leq\frac{C}{r_n}}\underbrace{\left|\frac{A_{ij}-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right|}_{\leq\frac{1}{\rho_n}}\right)$$

$$\leq \max_{i,j,i\neq j}\frac{n-2}{n-1}\left|\frac{1}{n-2}\sum_{\substack{t=1\\t\neq i,j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right|+\frac{C}{(n-1)r_n\rho_n}.$$

The expression inside the sum in the first term is bounded by $\frac{C}{r_n\rho_n}$, hence after conditioning on ξ_i, ξ_j we can apply the Bernstein's inequality for bounded i.i.d. random variables. The conditional expectation of that term is zero:

$$E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\middle|\xi_i,\xi_j\right)$$

$$=E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{1}{\rho_n}E\left(A_{tj}|\,\xi_i,\xi_j,\xi_t\right)-w_0(\xi_t,\xi_j)\right)\middle|\xi_i,\xi_j\right)$$

$$=E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(w_0(\xi_t,\xi_j)-w_0(\xi_t,\xi_j)\right)\middle|\xi_i,\xi_j\right)=0$$

where the first equality is due to the law of iterated expectations, the second uses the fact that d_{it}^2, $r_n(i)$ and $h_{0,n}(\xi_t,\xi_j)$ are not random after conditioning on ξ_i,ξ_j,ξ_t. A_{tj} is independent of ξ_i, hence $E\left(A_{tj}|\,\xi_i,\xi_j,\xi_t\right)=E\left(A_{tj}|\,\xi_j,\xi_t\right)$ which by definition equals $h_{0,n}(\xi_t,\xi_j)=\rho_n w_0(\xi_t,\xi_j)$.

The conditional variance is $O\left(\frac{1}{r_n\rho_n}\right)$:

$$\begin{aligned}
&Var\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\middle|\xi_i,\xi_j\right)\\
&=E\left(\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\right)^2 E\left(\left(\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)^2\middle|\xi_i,\xi_j,\xi_t\right)\middle|\xi_i,\xi_j\right)\\
&=E\left(\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\right)^2\left(w_0(\xi_t,\xi_j)\left(\frac{1}{\rho_n}-w_0(\xi_t,\xi_j)\right)\right)\middle|\xi_i,\xi_j\right)\\
&\le\frac{M_w}{\rho_n}\underbrace{\frac{E\left(\left(K\left(\frac{d_{it}^2}{b_n}\right)\right)^2\middle|\xi_i\right)}{r_n(i)^2}}_{=O\left(\frac{1}{r_n}\right)}=O\left(\frac{1}{r_n\rho_n}\right)
\end{aligned}$$

where in the last line we use that the kernel function is bounded ($K(\cdot)\le C$ by Assumption 1.3) and hence

$$\frac{E\left(\left(K\left(\frac{d_{it}^2}{b_n}\right)\right)^2\middle|\xi_i\right)}{r_n(i)^2}\le\frac{CE\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{\left(E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)\right)^2}=\frac{C}{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}\le\frac{C}{r_n}=O\left(\frac{1}{r_n}\right).$$

By union bound and Bernstein's inequality, for any $\varepsilon>0$ and $n\ge3$:

$$\begin{aligned}
&P\left(\max_{i,j,i\neq j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right|>\varepsilon\right)\\
&\le n(n-1)E\left(P\left(\left|\frac{1}{n-2}\sum_{\substack{t=1\\t\neq i,j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right|>\varepsilon-\frac{C}{(n-2)\rho_nr_n}\middle|\xi_i,\xi_j\right)\right)\\
&\le 2n(n-1)\exp\left(\frac{-(n-2)\left(\varepsilon-\frac{C}{(n-2)\rho_nr_n}\right)^2}{2\left(O\left(\frac{1}{r_n\rho_n}\right)+\frac{C}{3r_n\rho_n}\left(\varepsilon-\frac{C}{(n-2)\rho_nr_n}\right)\right)}\right)\\
&\le n^2\exp\left(-nr_n\rho_nC_\varepsilon\right)
\end{aligned}$$

for some $C_\varepsilon>0$. We can proceed in a very similar way for the case of $i=j$ to get:

$$P\left(\max_i\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq i}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\frac{A_{ti}-h_{0,n}(\xi_t,\xi_i)}{\rho_n}\right|>\varepsilon\right)\le O\left(n\exp\left(-nr_n\rho_nC_\varepsilon\right)\right).$$

Combining all the terms gives the required result: for any $\varepsilon>0$

$$\sum_{n=3}^{\infty} P\left(\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right|>\varepsilon\right)$$
$$\leq O\left(\sum_{n=3}^{\infty} n^2\exp\left(-nr_n\rho_n C_\varepsilon\right)\right)<\infty$$

under Assumption 1.1 and Assumption 1.4 which, by derivation similar to that at the end of the proof of 1.A.1, give:

$$\frac{n\rho_n r_n}{\log(n)}\geq C\frac{n\rho_n b_n^{\frac{1}{2\alpha}}}{\log(n)}\to\infty.$$

Hence

$$\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{A_{tj}-h_{0,n}(\xi_t,\xi_j)}{\rho_n}\right)\right|\xrightarrow{a.s.}0.$$

□

Lemma 1.A.4. *Under the assumptions of Theorem 1, for any $\varepsilon>0$:*

$$P\left(\left|\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{h_{0,n}(\xi_t,\xi_j)-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)\right|\right|>\varepsilon\right)$$
$$\leq O\left(n^2\exp\left(-nr_nC_\varepsilon\right)\right)+O\left(b_n^{\frac{\alpha^2}{(2\alpha+1)^2}}\right)\to 0.$$

hence

$$\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(\frac{h_{0,n}(\xi_t,\xi_j)-h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right)\right|\xrightarrow{p}0.$$

Proof. Intuitively, this result holds because as n increases $\frac{d_{it}^2}{b_n}$ becomes large, and hence $K\left(\frac{d_{it}^2}{b_n}\right)$ becomes zero, unless ξ_i and ξ_t are very close to each other in the sense that their $h_{0,n}(\xi_t,\xi_j)$ and $h_{0,n}(\xi_i,\xi_j)$ are similar for all ξ_j.

We start by showing that whenever $h_{0,n}(\xi_t,\xi_j)$ and $h_{0,n}(\xi_i,\xi_j)$ are not close, their distance d_{it} will be separated away from zero.

We follow the ideas from Auerbach (2022)'s proof of Lemma 1 which shows that for any i,t,n and any $\varepsilon>0$ we can find a $\delta>0$ such that $\sqrt{E\left(\left(w_0(\xi_i,\xi_j)-w_0(\xi_t,\xi_j)\right)^2\middle|\xi_i,\xi_t\right)}\geq$

$\varepsilon \implies d_{it} = \sqrt{E\left((\varphi(\xi_i, \xi_j) - \varphi(\xi_t, \xi_j))^2 \middle| \xi_i, \xi_t \right)} \geq \delta.$

Our idea is to add an extra step at the beginning: if for given i, j, t there is a $\nu > 0$ for which we have $|w_0(\xi_t, \xi_j) - w_0(\xi_i, \xi_j)| \equiv \left| \frac{h_{0,n}(\xi_t,\xi_j) - h_{0,n}(\xi_i,\xi_j)}{\rho_n} \right| > \nu$, then there exists an $\varepsilon > 0$ such that $\sqrt{E\left((w_0(\xi_i, \xi_j) - w_0(\xi_t, \xi_j))^2 \middle| \xi_i, \xi_t \right)} \geq \varepsilon$ (which in turn implies $d_{it} \geq \delta$). In other words, $|w_0(\xi_t, \xi_j) - w_0(\xi_i, \xi_j)|$ can be large only if d_{it} is large, in which case the weight placed on that term is small.

In our case, the issue is that in $\sqrt{E\left((w_0(\xi_i, \xi_j) - w_0(\xi_t, \xi_j))^2 \middle| \xi_i, \xi_t \right)}$ we take an expectation with respect to j, but the initial statement is given for a fixed j. To get around it, we replace the fixed j with a random element of a neighbourhood of j, then take an expectation with respect to an element of that neighbourhood, and use an upper bound which takes expectation over all possible values, not just those in the neighbourhood of j.

Recall from Assumption 1.2 that $N(\xi_j, \delta) = \left\{ \xi_k : \sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta \right\}$ denotes the neighbourhood of ξ_j of size δ. We fix i, j, t, k where $k \in N\left(\xi_j, \frac{\nu}{3}\right)$. Then

$$
\begin{aligned}
&\mathbb{1}\left(|w_0(\xi_t, \xi_j) - w_0(\xi_i, \xi_j)| > \nu\right) \\
&= \mathbb{1}\left(|w_0(\xi_t, \xi_j) - w_0(\xi_t, \xi_k) + w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k) + w_0(\xi_i, \xi_k) - w_0(\xi_i, \xi_j)| > \nu\right) \\
&\leq \underbrace{\mathbb{1}\left(|w_0(\xi_t, \xi_j) - w_0(\xi_t, \xi_k)| > \frac{\nu}{3}\right)}_{=0} + \mathbb{1}\left(|w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k)| > \frac{\nu}{3}\right) \\
&\quad + \underbrace{\mathbb{1}\left(|w_0(\xi_i, \xi_k) - w_0(\xi_i, \xi_j)| > \frac{\nu}{3}\right)}_{=0} \\
&= \mathbb{1}\left((w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k))^2 > \frac{\nu^2}{9} \right).
\end{aligned}
$$

If the above holds for any fixed i, j, t and for any $\xi_k \in N\left(\xi_j, \frac{\nu}{3}\right)$, it also holds if we take expectation over $\xi_k \in N\left(\xi_j, \frac{\nu}{3}\right)$. Recall from Assumption 1.2 that $\omega(\delta) = \inf_{\xi_j} P\left(\xi_k \in N(\xi_j, \delta) \middle| \xi_j\right)$ and $\omega(\delta) \geq \left(\frac{\delta}{C}\right)^{\frac{1}{\alpha}}$ for all $\delta > 0$. We use $E(X|A) = \frac{E(\mathbb{1}_A X)}{P(A)}$.

$$
\begin{aligned}
&\mathbb{1}\left(|w_0(\xi_t, \xi_j) - w_0(\xi_i, \xi_j)| > \nu\right) \\
&\leq \mathbb{1}\left(E\left((w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k))^2 \middle| \xi_k \in N\left(\xi_j, \frac{\nu}{3}\right), \xi_i, \xi_j, \xi_t \right) > \frac{\nu^2}{9} \right) \\
&= \mathbb{1}\left(\frac{E\left((w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k))^2 \mathbb{1}\left(\xi_k \in N\left(\xi_j, \frac{\nu}{3}\right)\right) \middle| \xi_i, \xi_j, \xi_t \right)}{P\left(\xi_k \in N\left(\xi_j, \frac{\nu}{3}\right) \middle| \xi_j\right)} > \frac{\nu^2}{9} \right) \\
&\leq \mathbb{1}\left(E\left((w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k))^2 \mathbb{1}\left(\xi_k \in N\left(\xi_j, \frac{\nu}{3}\right)\right) \middle| \xi_i, \xi_j, \xi_t \right) > \frac{\nu^2}{9} \omega\left(\frac{\nu}{3}\right) \right) \\
&\leq \mathbb{1}\left(E\left((w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k))^2 \middle| \xi_i, \xi_t \right) > \frac{\nu^2}{9} \omega\left(\frac{\nu}{3}\right) \right) \\
&= \mathbb{1}\left(\sqrt{E\left((w_0(\xi_t, \xi_k) - w_0(\xi_i, \xi_k))^2 \middle| \xi_i, \xi_t \right)} > \frac{\nu}{3} \sqrt{\omega\left(\frac{\nu}{3}\right)} \right)
\end{aligned}
$$

We set $\varepsilon = \frac{\nu}{3}\sqrt{\omega\left(\frac{\nu}{3}\right)}$, which completes the argument.

Like Auerbach (2022),[13] we assume there exist some $\alpha, C > 0$ such that for any δ we have $\omega(\delta) \geq \left(\frac{\delta}{C}\right)^{\frac{1}{\alpha}}$ (this is our Assumption 1.2). Then:

$$\begin{aligned}\mathbb{1}\left(|w_0(\xi_t,\xi_j) - w_0(\xi_i,\xi_j)| > \nu\right) &\leq \mathbb{1}\left(\sqrt{E\left(\left(w_0(\xi_t,\xi_k) - w_0(\xi_i,\xi_k)\right)^2 \middle| \xi_i,\xi_t\right)} > \frac{\nu}{3}\sqrt{\left(\frac{\nu}{3C}\right)^{\frac{1}{\alpha}}}\right) \\ &= \mathbb{1}\left(3C^{\frac{1}{2\alpha+1}}\left(E\left(\left(w_0(\xi_t,\xi_k) - w_0(\xi_i,\xi_k)\right)^2 \middle| \xi_i,\xi_t\right)\right)^{\frac{\alpha}{2\alpha+1}} > \nu\right).\end{aligned}$$

hence

$$|w_0(\xi_t,\xi_j) - w_0(\xi_i,\xi_j)| \leq 3C^{\frac{1}{2\alpha+1}}\left(E\left(\left(w_0(\xi_t,\xi_k) - w_0(\xi_i,\xi_k)\right)^2 \middle| \xi_i,\xi_t\right)\right)^{\frac{\alpha}{2\alpha+1}}.$$

Combining with Auerbach (2022) result:[14]

$$\sqrt{E\left(\left(w_0(\xi_t,\xi_k) - w_0(\xi_i,\xi_k)\right)^2 \middle| \xi_i,\xi_t\right)} \leq 2C^{\frac{1}{4\alpha+2}} d_{it}^{\frac{\alpha}{2\alpha+1}}.$$

we get

$$|w_0(\xi_t,\xi_j) - w_0(\xi_i,\xi_j)| \leq \tilde{C} d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}$$

for $\tilde{C} = 3 \times 2^{\frac{2\alpha}{2\alpha+1}} \times C^{\frac{3\alpha+1}{(2\alpha+1)^2}}$.

We can now return to the term of interest.

$$\begin{aligned}\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\left(w_0(\xi_t,\xi_j) - w_0(\xi_i,\xi_j)\right)\right| &\leq \max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\tilde{C} d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}\right| \\ \leq \tilde{C}\left(\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)} d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}} - E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)} d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}\middle| \xi_i\right)\right| + \right. \\ \left. + \max_{i}\left|E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)} d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}\middle| \xi_i\right)\right|\right).\end{aligned}$$

The first term goes to zero by the union bound and Bernstein's inequality, where, conditionally on ξ_i, ξ_j and after separating out the term with $t = i$, the terms inside the average are i.i.d.,

[13] One major difference is that Auerbach (2022) does not allow for sparsity in his model, in his case $\rho_n = 1$. Hence we impose an assumption analogous to his to w_0, not $h_{0,n}$. If we were to define everything in terms of $h_{0,n}$, we would need $N_n(\xi_j,\delta) = \left\{\xi_k : \sup_{\xi_t} |h_{0,n}(\xi_t,\xi_k) - h_{0,n}(\xi_t,\xi_j)| < \delta\right\}$, $\omega_n(\delta) = \inf_{\xi_j} P\left(\xi_k \in N_n(\xi_j,\delta) | \xi_j\right)$ and $\omega_n(\delta) \geq \left(\frac{\delta}{\rho_n C}\right)^{\frac{1}{\alpha}}$.

[14] This is Auerbach (2022) Lemma A1 restated in our notation. To account for the fact that Auerbach (2022) does not allow for sparsity we replace their f, which is equivalent to our $h_{0,n}$, with a w_0 and their δ with our equivalent term d.

mean zero, bounded by $r_n^{-1}CM_w^{\frac{4\alpha^2}{(2\alpha+1)^2}} = O\left(r_n^{-1}\right)$ and have variance $O\left(r_n^{-1}\right)$. For any $\varepsilon > 0$:

$$
\begin{aligned}
&P\left(\max_{i,j}\left|\frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n}\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}} - E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}\middle|\xi_i\right)\right| > \varepsilon\right)\\
&\leq 2n(n-1)\exp\left(-\frac{(n-2)\left(\varepsilon - \frac{CM_w^{\frac{4\alpha^2}{(2\alpha+1)^2}}}{(n-2)r_n}\right)^2}{2\left(O\left(r_n^{-1}\right) + \frac{CM_w^{\frac{4\alpha^2}{(2\alpha+1)^2}}}{3r_n}\left(\varepsilon - \frac{CM_w^{\frac{4\alpha^2}{(2\alpha+1)^2}}}{(n-2)r_n}\right)\right)}\right)\\
&+ 2n\exp\left(-\frac{(n-1)\varepsilon^2}{2\left(O\left(r_n^{-1}\right) + \frac{CM_w^{\frac{4\alpha^2}{(2\alpha+1)^2}}\varepsilon}{3r_n}\right)}\right)\\
&\leq n^2\exp\left(-nr_nC_\varepsilon\right) \to 0.
\end{aligned}
$$

The last convergence was shown at the end of the proof of Lemma 1.A.1.

It remains to show that the last term goes to zero too. By Assumption 1.3, there exists a $D \in \mathbb{R}$ such that $\forall |u| > D : K(u) = 0$. If $d_{it} \neq 0$: $\frac{d_{it}^2}{b_n} = O\left(\frac{1}{b_n}\right) \to \infty$, so eventually, as $n \to \infty$, $\frac{d_{it}^2}{b_n} > D$ and $K\left(\frac{d_{it}^2}{b_n}\right) = 0$ (and if $d_{it} = 0$ the whole term is identically equal to zero). We have:

$$
\begin{aligned}
\max_i\left|E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}\middle|\xi_i\right)\right| &= \max_i\left|E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}d_{it}^{\frac{2\alpha^2}{(2\alpha+1)^2}}\mathbb{1}\left(\frac{d_{it}^2}{b_n} \leq D\right)\middle|\xi_i\right)\right|\\
&\leq \max_i\left|(Db_n)^{\frac{\alpha^2}{(2\alpha+1)^2}}E\left(\frac{K\left(\frac{d_{it}^2}{b_n}\right)}{r_n(i)}\middle|\xi_i\right)\right|\\
&= \max_i\left|(Db_n)^{\frac{\alpha^2}{(2\alpha+1)^2}}\underbrace{\frac{E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\xi_i\right)}{r_n(i)}}_{=1}\right|\\
&\leq D^{\frac{\alpha^2}{(2\alpha+1)^2}}b_n^{\frac{\alpha^2}{(2\alpha+1)^2}} = O\left(b_n^{\frac{\alpha^2}{(2\alpha+1)^2}}\right) \to 0.
\end{aligned}
$$

The last expression goes to zero by Assumption 1.4. Note however that under Assumption 1.4 the rate of convergence to zero is too slow to ensure almost sure convergence of this term. This is the reason why we only get uniform convergence in probability in Theorem 1 and convergence weakly in probability in Theorem 2.4.2.

□

Lemma 1.A.5. *Under the assumptions of Theorem 1, for any $\varepsilon > 0$:*

$$\sum_{n=3}^{\infty} P\left(\frac{1}{b_n r_n}\max_{i,j}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| > \varepsilon\right) = O\left(\sum_{n=3}^{\infty} n^2 \exp\left(-nb_n^2 r_n^2 \rho_n^2 C_\varepsilon\right)\right) = O(1)$$

hence

$$\frac{1}{b_n r_n}\max_{i,j}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| \xrightarrow{a.s.} 0.$$

Proof. We follow the same steps as in Lemma B1 in Auerbach (2022). By definition:

$$\begin{aligned}
\hat{d}_{ij} &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\sum_{s=1}^{n}\frac{A_{ts}}{\rho_n}\left(\frac{A_{is}-A_{js}}{\rho_n}\right)\right)^2} \\
\tilde{d}_{ij} &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\varphi(\xi_i,\xi_t)-\varphi(\xi_j,\xi_t)\right)^2} \\
d_{ij} &= \sqrt{E_t\left(\left(\varphi(\xi_i,\xi_t)-\varphi(\xi_j,\xi_t)\right)^2 \middle| \xi_i,\xi_j\right)}.
\end{aligned}$$

Take any $\varepsilon > 0$. We have:

$$\begin{aligned}
&P\left(\frac{1}{b_n r_n}\max_{i,j}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| > \varepsilon\right) = P\left(\max_{i,j}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| > \varepsilon b_n r_n\right) \\
&= P\left(\max_{i,j}\left|\hat{d}_{ij}^2 - \tilde{d}_{ij}^2 + \tilde{d}_{ij}^2 - d_{ij}^2\right| > \varepsilon b_n r_n\right) \\
&\leq P\left(\max_{i,j}\left|\hat{d}_{ij}^2 - \tilde{d}_{ij}^2\right| > \frac{\varepsilon b_n r_n}{2}\right) + P\left(\max_{i,j}\left|\tilde{d}_{ij}^2 - d_{ij}^2\right| > \frac{\varepsilon b_n r_n}{2}\right) \qquad (1.15)
\end{aligned}$$

where the last inequality follows from the fact that $|a+b| > \varepsilon$ implies $|a| > \frac{\varepsilon}{2}$ or $|b| > \frac{\varepsilon}{2}$ and hence $P(|a+b| > \varepsilon) \leq P(|a| > \frac{\varepsilon}{2}) + P(|b| > \frac{\varepsilon}{2})$.

For the first term in (1.15), we plug in the definitions, then use $a^2 - b^2 = (a-b)(a+b)$ and the fact that the second bracket approaches a limit bounded by $4M_w^2$:

$$\begin{aligned}
&P\left(\max_{i,j}\left|\hat{d}_{ij}^2-\tilde{d}_{ij}^2\right|>\frac{\varepsilon b_n r_n}{2}\right)\\
&=P\left(\max_{i,j}\left|\frac{1}{n}\sum_{t=1}^n\left(\left(\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}}{\rho_n}\left(\frac{A_{is}}{\rho_n}-\frac{A_{js}}{\rho_n}\right)\right)^2-(\varphi(\xi_i,\xi_t)-\varphi(\xi_j,\xi_t))^2\right)\right|>\frac{\varepsilon b_n r_n}{2}\right)\\
&=P\left(\max_{i,j}\left|\frac{1}{n}\sum_{t=1}^n\left(\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}-\varphi(\xi_i,\xi_t)+\varphi(\xi_j,\xi_t)-\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{js}}{\rho_n^2}\right)\times\right.\right.\\
&\qquad\left.\left.\times\left(\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}+\varphi(\xi_i,\xi_t)-\varphi(\xi_j,\xi_t)-\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{js}}{\rho_n^2}\right)\right|>\frac{\varepsilon b_n r_n}{2}\right)\\
&\leq P\left(\max_{i,j}\left|\frac{1}{n}\sum_{t=1}^n\left(\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}-\varphi(\xi_i,\xi_t)+\varphi(\xi_j,\xi_t)-\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{js}}{\rho_n^2}\right)\right|>\frac{\varepsilon b_n r_n}{16M_w^2}\right)+\\
&\qquad+P\left(\max_{i,j,t}\left|\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}+\varphi(\xi_i,\xi_t)-\varphi(\xi_j,\xi_t)-\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{js}}{\rho_n^2}\right|>8M_w^2\right)
\end{aligned}$$

where the last equality follows from the fact that $ab>\varepsilon$ implies $b\geq M$ or $a>\frac{\varepsilon}{M}$. For the first term, we again note that $|a+b|>\varepsilon$ implies $|a|>\frac{\varepsilon}{2}$ or $|b|>\frac{\varepsilon}{2}$, we split the expression into a part with terms that only depend on i and a part with terms that only depend on j. We then get:

$$\begin{aligned}
&P\left(\max_{i,j}\left|\frac{1}{n}\sum_{t=1}^n\left(\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}-\varphi(\xi_i,\xi_t)+\varphi(\xi_j,\xi_t)-\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{js}}{\rho_n^2}\right)\right|>\frac{\varepsilon b_n r_n}{16M_w^2}\right)\\
&\leq 2P\left(\max_{i}\left|\frac{1}{n}\sum_{t=1}^n\left(\frac{1}{n}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}-\varphi(\xi_i,\xi_t)\right)\right|>\frac{\varepsilon b_n r_n}{32M_w^2}\right)\\
&\leq 2nE\left(P\left(\max_{t,t\neq i}\left|\frac{1}{n-1}\sum_{s=1}^n\frac{A_{ts}A_{is}}{\rho_n^2}-\varphi(\xi_i,\xi_t)\right|>\frac{\varepsilon b_n r_n}{32M_w^2}-\frac{1}{(n-1)\rho_n^2}\middle|\xi_i\right)\right)\\
&\leq 2n(n-1)\\
&\qquad E\left(P\left(\left|\frac{1}{n-2}\sum_{\substack{s=1\\s\neq i,t}}^n\frac{A_{ts}A_{is}}{\rho_n^2}-\varphi(\xi_i,\xi_t)\right|>\frac{\varepsilon b_n r_n}{32M_w^2}-\frac{1}{(n-1)\rho_n^2}-\frac{2}{(n-2)\rho_n^2}\middle|\xi_i,\xi_t\right)\right)\\
&\leq 4n(n-1)\exp\left(\frac{-(n-2)\left(\frac{\varepsilon b_n r_n}{32M_w^2}-\frac{1}{(n-1)\rho_n^2}-\frac{2}{(n-2)\rho_n^2}\right)^2}{2\left(\frac{C}{\rho_n^2}+\frac{1}{3}\frac{1}{\rho_n^2}\left(\frac{\varepsilon b_n r_n}{32M_w^2}-\frac{1}{(n-1)\rho_n^2}-\frac{2}{(n-2)\rho_n^2}\right)\right)}\right)\\
&\leq n^2\exp\left(-nb_n^2r_n^2\rho_n^2C_\varepsilon\right)
\end{aligned}$$

where the second inequality follows from the union bound applied to $\max_i$ and the fact that $\frac{1}{n}\sum_{t=1}^n x_t\leq\frac{n-1}{n}\frac{1}{n-1}\sum_{\substack{t=1\\t\neq i}}^n\max_{t,t\neq i}x_t+\frac{1}{n}x_i=\frac{n-1}{n}\left(\max_{t,t\neq i}x_t+\frac{1}{n-1}x_i\right)$. In this case $x_i=\frac{1}{n}\sum_{s=1}^n\frac{A_{is}^2}{\rho_n^2}-\varphi(\xi_i,\xi_i)$ and $|x_i|\leq\frac{1}{\rho_n^2}$ (since A and $\rho_n^2\varphi$ both belong to $[0,1]$). We also use the fact that $\frac{n-1}{n}|a|>\varepsilon$ implies $|a|>\varepsilon$. Next, notice that $\left|a\pm\frac{1}{(n-1)\rho_n^2}\right|>\varepsilon$ implies that either $|a|\geq a>\varepsilon\pm\frac{1}{(n-1)\rho_n^2}>\varepsilon-\frac{1}{(n-1)\rho_n^2}$ or $|a|\geq -a>\varepsilon\pm\frac{1}{(n-1)\rho_n^2}>\varepsilon-\frac{1}{(n-1)\rho_n^2}$, so in either case we get $|a| > \varepsilon - \frac{1}{(n-1)\rho_n^2}$. For the third inequality, we again apply the union bound, this time over $t \neq i$, and separate out the terms with $s = i$ or $s = t$, similarly to the previous step. The final inequality follows from Bernstein's inequality with $\left|\frac{A_{ts}A_{is}}{\rho_n^2} - \varphi(\xi_i, \xi_t)\right| \leq \frac{1}{\rho_n^2}$ and $Var\left(\frac{A_{ts}A_{is}}{\rho_n^2} - \varphi(\xi_i, \xi_t)\right) = E\left(\varphi\left(\xi_i, \xi_t\right)\left(\frac{1}{\rho_n^2} - \varphi\left(\xi_i, \xi_t\right)\right)\right) \leq \frac{M_w^2}{\rho_n^2}$.

For the second term, take any $0 < \varepsilon < 2M_w^2$ and use similar arguments.

$$
\begin{aligned}
&P\left(\max_{i,j,t}\left|\frac{1}{n}\sum_{s=1}^{n}\frac{A_{ts}A_{is}}{\rho_n^2} + \varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t) - \frac{1}{n}\sum_{s=1}^{n}\frac{A_{ts}A_{js}}{\rho_n^2}\right| > 8M_w^2\right)\\
&\leq 2P\left(\max_{i,t}\left|\frac{1}{n}\sum_{s=1}^{n}\frac{A_{ts}A_{is}}{\rho_n^2} + \varphi(\xi_i, \xi_t)\right| > 4M_w^2\right)\\
&\leq 2\Bigg(\underbrace{P\left(\max_{i,t}\left|\frac{1}{n}\sum_{s=1}^{n}\frac{A_{ts}A_{is}}{\rho_n^2} - \varphi(\xi_i, \xi_t)\right| > \varepsilon\right)}_{\leq 2n^2\exp\left(\frac{-(n-2)\left(\varepsilon - \frac{2}{(n-1)\rho_n^2}\right)^2}{2\left(O\left(\frac{1}{\rho_n^2}\right) + \frac{1}{3}\frac{1}{\rho_n^2}\left(\varepsilon - \frac{2}{(n-1)\rho_n^2}\right)\right)}\right)} + \underbrace{P\Bigg(\underbrace{\max_{i,t}|2\varphi(\xi_i, \xi_t)|}_{\leq 2M_w^2} > \underbrace{4M_w^2 - \varepsilon}_{>2M_w^2}\Bigg)}_{=0}\Bigg)\\
&\leq n^2\exp\left(-n\rho_n^2 C_\varepsilon\right)
\end{aligned}
$$

We show that the second term in (1.15) goes to zero almost surely by applying the union bound and Bernstein's inequality. We also use that, by definition, $d_{ii} = \tilde{d}_{ii} = 0$.

$$
\begin{aligned}
&P\left(\max_{i,j}\left|\tilde{d}_{ij}^2 - d_{ij}^2\right| > \frac{\varepsilon b_n r_n}{2}\right)\\
&\leq P\left(\max_{i,j,i\neq j}\left|\tilde{d}_{ij}^2 - d_{ij}^2\right| > \frac{\varepsilon b_n r_n}{2}\right) + \underbrace{P\left(\underbrace{\max_{i}\left|\tilde{d}_{ii}^2 - d_{ii}^2\right|}_{=0} > \frac{\varepsilon b_n r_n}{2}\right)}_{=0}\\
&\leq n(n-1)E\left(P\left(\left|\tilde{d}_{ij}^2 - d_{ij}^2\right| > \frac{\varepsilon b_n r_n}{2}\middle|\xi_i, \xi_j\right)\right)\\
&= n(n-1)E\Bigg(P\Bigg(\left|\frac{1}{n}\sum_{t=1}^{n}\left((\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t))^2 - E\left((\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t))^2\middle|\xi_i, \xi_j\right)\right)\right|\\
&\qquad\qquad > \frac{\varepsilon b_n r_n}{2}\Bigg|\xi_i, \xi_j\Bigg)\Bigg)\\
&\leq n(n-1)E\Bigg(P\Bigg(\Bigg|\frac{1}{n-2}\sum_{\substack{t=1\\ t\neq i,j}}^{n}\Big((\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t))^2\\
&\qquad\qquad - E\left((\varphi(\xi_i, \xi_t) - \varphi(\xi_j, \xi_t))^2\middle|\xi_i, \xi_j\right)\Big)\Bigg| > \frac{\varepsilon b_n r_n}{2} - \frac{2}{(n-2)\rho_n^2}\Bigg|\xi_i, \xi_j\Bigg)\Bigg)\\
&\leq 2n(n-1)\exp\left(\frac{-(n-2)\left(\frac{\varepsilon b_n r_n}{2} - \frac{2}{(n-2)\rho_n^2}\right)^2}{2 + \frac{2\left(\frac{\varepsilon b_n r_n}{2} - \frac{2}{(n-2)\rho_n^2}\right)}{3}}\right) \leq n^2\exp\left(-n b_n^2 r_n^2 C_\varepsilon\right).
\end{aligned}
$$

The conclusion follows since

$$\sum_{n=3}^{\infty} P\left(\frac{1}{b_n r_n} \max_{i,j} \left| \hat{d}_{ij}^2 - d_{ij}^2 \right| > \varepsilon \right) = O\left(\sum_{n=3}^{\infty} n^2 \exp\left(-n b_n^2 r_n^2 \rho_n^2 C_\varepsilon\right)\right) = O(1)$$

The last term is bounded for any $C_\varepsilon > 0$ because under Assumption 1.4 $\frac{\log(n)}{n b_n^2 r_n^2 \rho_n^2} \to 0$. □

Subsection 1.A.2 Useful results

For reference, we list some results which we use in our proofs:

Theorem (Bernstein's inequality for bounded random variables[15])**.** *Let $Z_1, \dots, Z_n$ be independent random variables. Assume that there exist some positive constant M such that $|Z_i| \leq M$ with probability one for each i. Let also $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} V(Z_i)$. Then, for all $\varepsilon > 0$:*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E(Z_i))\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2\left(\sigma^2 + \frac{1}{3}M\varepsilon\right)}\right). \tag{1.16}$$

Appendix 1.B Additional tables, codes

Subsection 1.B.1 Tables

[15] Copied after Zeleneev (2020).

			observed	HK1 for $c =$							HK2	HNN1	dot product for $k =$	
				0.01	0.1	0.5	1	2	10	100			1	5
n	function	ρ_n												
100	high density	0.45	0.483	0.483	0.483	0.483	0.483	0.483	0.066	0.010	0.484	0.021	0.015	0.073
		0.59	0.462	0.462	0.462	0.462	0.462	0.461	0.063	0.014	0.461	0.023	0.018	0.072
		0.76	0.342	0.342	0.342	0.274	0.074	0.024	0.021	0.022	0.048	0.017	0.023	0.061
	horseshoe	0.07	0.102	0.102	0.102	0.093	0.054	0.016	0.021	0.023	0.043	0.017	0.023	0.025
		0.09	0.121	0.121	0.121	0.107	0.052	0.019	0.035	0.038	0.042	0.022	0.037	0.034
		0.11	0.134	0.134	0.134	0.114	0.048	0.023	0.056	0.063	0.040	0.028	0.060	0.047
	product	0.15	0.221	0.221	0.219	0.178	0.083	0.013	0.009	0.014	0.067	0.009	0.006	0.035
		0.19	0.256	0.256	0.253	0.195	0.064	0.010	0.012	0.023	0.071	0.011	0.006	0.041
		0.25	0.281	0.281	0.277	0.136	0.018	0.009	0.015	0.038	0.061	0.012	0.006	0.042
500	high density	0.24	0.357	0.357	0.357	0.357	0.357	0.357	0.003	0.002	0.357	0.005	0.003	0.014
		0.42	0.477	0.477	0.477	0.477	0.477	0.477	0.111	0.006	0.477	0.007	0.007	0.021
		0.76	0.338	0.338	0.338	0.317	0.145	0.008	0.009	0.018	0.030	0.004	0.018	0.028
	horseshoe	0.04	0.062	0.062	0.062	0.060	0.037	0.006	0.006	0.006	0.030	0.003	0.006	0.005
		0.06	0.098	0.098	0.098	0.093	0.042	0.005	0.015	0.019	0.032	0.004	0.018	0.012
		0.11	0.136	0.136	0.136	0.084	0.009	0.005	0.035	0.061	0.010	0.007	0.056	0.033
	product	0.08	0.134	0.134	0.132	0.108	0.063	0.005	0.002	0.003	0.057	0.002	0.001	0.006
		0.14	0.209	0.209	0.208	0.178	0.114	0.006	0.002	0.010	0.097	0.003	0.001	0.009
		0.25	0.279	0.279	0.278	0.247	0.141	0.005	0.002	0.011	0.060	0.004	0.001	0.011
1000	high density	0.18	0.287	0.287	0.287	0.287	0.287	0.025	0.001	0.001	0.286		0.001	0.006
		0.37	0.455	0.455	0.455	0.455	0.455	0.455	0.148	0.004	0.455		0.005	0.012
		0.76	0.335	0.335	0.335	0.320	0.176	0.007	0.002	0.016	0.015		0.017	0.022
	horseshoe	0.03	0.047	0.047	0.047	0.045	0.028	0.005	0.003	0.003	0.024		0.003	0.002
		0.05	0.087	0.087	0.087	0.084	0.040	0.003	0.010	0.014	0.031		0.013	0.008
		0.11	0.137	0.137	0.136	0.027	0.003	0.003	0.027	0.060	0.005		0.056	0.032
	product	0.06	0.104	0.104	0.102	0.086	0.057	0.007	0.001	0.002	0.049		0.000	0.002
		0.12	0.188	0.188	0.187	0.160	0.105	0.006	0.001	0.007	0.089		0.001	0.004
		0.25	0.277	0.277	0.276	0.249	0.162	0.004	0.001	0.003	0.030		0.001	0.005

Table 1.1: MSE for simulated fits for different true generated functions at different density levels and sample sizes. Observed is the binary adjacency matrix, HK1 is our method with different bandwidths of the form $c \times \hat{a}$, HK2 is our alternative specification. HNN1 is the nearest neighbours method of Zhang, Levina, and Zhu (2017). Some values are missing due to too high memory requirements. dot product is the method from Levin and Levina (2019) with k-dimensional ξ_i.

Subsection 1.B.2 Codes

In this section we present some of the codes used for simulations. A full package should eventually become available online.

We start with the definitions of different distances and estimators of the linking functions:

```
def D2(A):
    #a function which maps A into a matrix of distances D2
    n= len(A)
    V=(1/n)*np.matmul(A,A)
    C=(1/n)*np.matmul(V,V)
    B=np.matmul(np.diag(np.diag(C)),np.ones((n,n)))
    D = B+B.T-2*C
    return D
```

```
def Dmax(A):
    #a function which maps A into a matrix of distances Dmax
    n = len(A)
    V = (1/n)*np.matmul(A,A)
    F = torch.tensor(np.tensordot(np.ones(n),V,0))
    G = torch.transpose(F, 1,0)
    H = F-G
    J=1-torch.transpose(np.fmax(torch.eye(n).repeat(n, 1, 1),np.tensordot(np.eye
                                        (n),np.ones(n),0)),2,1)
    D = np.array(torch.amax(np.fmin(abs(H),J), dim=2))
    return D
```

```
def HK1h(D,A,h):
    # gives a kernel approximation to the linking function based on a one-way
                                        normal kernel, with bandwidth h,
                                        based on distance D
    n= len(A)
    K = np.exp(-0.5*(D/h)**2)
    K[np.isnan(K)] = 0
    T=np.matmul(K,A)
    B=np.matmul(K,np.ones((n,n)))-K
    H = T/B
    H=(H+np.transpose(H))/2
    return H
```

```
def HK2h(D,A,h):
    # gives a kernel approximation to the linking function based on a two-way
                                        normal kernel, with bandwidth h,
                                        based on distance D
    n= len(A)
    K = np.exp(-0.5*(D/h)**2)
    K[np.isnan(K)] = 0
    H=np.matmul(K,np.matmul(A,K))/np.matmul(K,np.matmul(1-np.eye(n),K))
    return H
```

```
def HNN1(D,A):
    #gives a kernel approximation to h based on one-way nearest neighbours, with
                                                 bandwidth h, based on distance D
    #uses the optimal neighbourhood size (n log(n))^{1/2}
    n= len(A)
    N_size = round(np.sqrt(n*np.log(n)))
    N=np.argpartition(D, N_size+1)[:,:N_size+1]
    mask = np.ones((n,N_size+1), dtype=bool)
    mask[range(n), np.argmax(N==np.array(range(n)).reshape(n,1), axis=1)] =
                                                False
    N = N[mask].reshape(n, N_size)
    mask2 = np.zeros((n,n), dtype=bool)
    mask2[np.tile(np.array(range(n)).reshape(n,1),N_size), N] = True
    mask_long=np.tile(mask2,(n,1))
    A_long=np.tile(A,(1,n)).reshape(n*n,n)
    Amlong = A_long[mask_long].reshape(n,n, N_size)
    H=np.sum(Amlong,2)/N_size
    H=(H+np.transpose(H))/2
    return H
```

In simulations we generate the true matrices using one of the following functions:

```
def high_rho_generate(n, r, rep):
    #generates a rep number of true n by n matrices from the high rho function
                                                with density r/1.35, outputs only
                                                the adjacency matrices
    A_true = []
    for s in range(rep):
        w = np.random.uniform(0,1,(n))
        u = np.random.uniform(0,1,(n,n))
        eta = np.tril(u) + np.tril(u, -1).T
        Wi = np.tensordot(w,np.ones(n),0)
        Wj = np.tensordot(np.ones(n),w,0)
        A = (eta < r*(1-((abs(0.5-Wi)<0.05) & (abs(0.5-Wj)<0.05)))*(1-0.5*(abs(0
                                                    .5-Wi)+abs(0.5-Wj)))/(0.975))*1
        np.fill_diagonal(A, 0)
        A_true.append(A)
    return A_true
```

```
def horse_generate(n, r, rep):
    #generates a rep number of true n by n matrices from the horseshoe function
                                             with density r/4.44, outputs only
                                             the adjacency matrices
    A_true = []
    for s in range(rep):
        w = np.random.uniform(0,1,(n))
        u = np.random.uniform(0,1,(n,n))
        eta = np.tril(u) + np.tril(u, -1).T
        Wi = np.tensordot(w,np.ones(n),0)
        Wj = np.tensordot(np.ones(n),w,0)
        A = (eta < r*((np.exp(-200*(Wi-Wj**2)**2)+np.exp(-200*(Wj-Wi**2)**2))/2)
                                                 )*1
        np.fill_diagonal(A, 0)
        A_true.append(A)
    return A_true
```

```
def product_generate_A_h_xi(n, r, rep):
    #generates a rep number of true n by n matrices from the product function
                                             with density r/4, outputs the
                                             adjacency matrices, the true linking
                                              function, and the true values of
                                             the underlying characteristics $\
                                             xi_i$
    A_true = []
    xi_true = []
    h_true = []
    for s in range(rep):
        w = np.random.uniform(0,1,(n))
        u = np.random.uniform(0,1,(n,n))
        eta = np.tril(u) + np.tril(u, -1).T
        Wi = np.tensordot(w,np.ones(n),0)
        Wj = np.tensordot(np.ones(n),w,0)
        h = r*Wi*Wj
        A = (eta < h)*1
        np.fill_diagonal(A, 0)
        np.fill_diagonal(h, 0)
        A_true.append(A)
        h_true.append(h)
        xi_true.append(list(w))
    return (A_true, h_true, np.array(xi_true))
```

Code for finding the optimal bandwidth:

```
    def HK1h_loo(D,A,h):
        #gives a leave-obe-out kernel approximation to h based on a one-way
```

```
                                                 normal kernel, with bandwidth h,
                                                  based on distance D
        n= len(A)
        K = np.exp(-0.5*(D/h)**2)
        K[np.isnan(K)] = 0
        T = np.matmul(K,A) - np.matmul(np.diag(np.diag(K)),np.ones((n,n)))*A
        B = np.matmul(K,np.ones((n,n)))-(K-np.diag(np.diag(K)))-np.matmul(np.
                                                 diag(np.diag(K)),np.ones((n,n)))
        H = T/B
        H=(H+np.transpose(H))/2
        return H
```

```
def log_likelihood(A,H):
    #the log-likelihood estimation for an adjacency matrix A under the
                                             assumption it comes from a
                                             distribution with linking
                                             probabilities in H
    log_likelihood = np.sum(A*np.log(H)+(1-A)*np.log(1-H))
    return log_likelihood
```

```
def ll(h, A):
    #the leave-one-out log-likelihood objective function for use in minimising
                                             procedures
    return -log_likelihood(A,HK1h_loo(D2(A),A,h))
```

Code for simulations:

```
    #simulations. loop over:
NN = [100, 200, 500, 1000]
rho_type = ['high', 'boundary','constant']
function = ['product', 'horseshoe', 'high density']
methods = ['true', 'observed', 'HK1', 'HK2','HK1_penalty', 'HNN1', 'dot_prod_1',
                                          'dot_prod_5']


#bandwidth choice: only for HK1
CC = [0.01, 0.1, 0.5, 1, 2, 10, 100]


#loop:
for n in NN:
    for rt in rho_type:
        if rt == 'constant':
            r=1
        elif rt == 'high':
            r=np.sqrt(np.sqrt((25/np.log(25))))*np.sqrt(np.sqrt(np.log(n)/n))
        elif rt == 'boundary':
```

```
        r=np.sqrt((25/np.log(25)))*np.sqrt(np.log(n)/n)
    for fun in function:
        np.random.seed(27)
        w=np.linspace(0,1,n)
        u = np.random.uniform(0,1,(n,n))
        eta = np.tril(u) + np.tril(u, -1).T
        Wi = np.tensordot(w,np.ones(n),0)
        Wj = np.tensordot(np.ones(n),w,0)
        if fun == ’product’:
            H = r*Wi*Wj
        elif fun == ’horseshoe’:
            H = r*((np.exp(-200*(Wi-Wj**2)**2)+np.exp(-200*(Wj-Wi**2)**2))/2
                                                     )
        elif fun == ’high density’:
            H = r*(1-((abs(0.5-Wi)<0.05) & (abs(0.5-Wj)<0.05)))*(1-0.5*(abs(
                                                     0.5-Wi)+abs(0.5-Wj)))/(0
                                                     .975)
        A = (eta < H)*1
        np.fill_diagonal(A, 0)

        for met in methods:
            if met == ’HK1’:
                h_guess = 0.2090189845643738*0.1**1.38258532*n**(-1.55268817
                                                         )*np.log(n)**1.
                                                         82661653
                res = minimize(ll, h_guess, args=A, method = ’Nelder-Mead’,
                                                         tol=1e-7, bounds=((0
                                                         ,1.1),))
                h = res.x[0]
                for c in CC:
                    Matrix = HK1h(D2(A),A,c*h)
                    np.savetxt(’matrices_simulation/’+’Matrix_’+’n_’+str(n)+
                                                             ’_’+’r_’+rt+’_’+
                                                             fun+’_’+met+’_c_
                                                             ’+str(c)+’.csv’,
                                                              Matrix,
                                                             delimiter=",")
            if met != ’HK1’:
                if met == ’true’:
                    Matrix = H
                elif met == ’observed’:
                    Matrix = A
                elif met == ’HK2’:
                    res = minimize(ll_HK2, h, args=A, method = ’Nelder-Mead’
                                                             , bounds=((0,1.1
                                                             ),))
                    h2 = res.x[0]
                    Matrix = HK2h(D2(A),A,h2)
                elif met == ’HNN1’:
                    Matrix = HNN1(Dmax(A),A)
                elif met == ’dot_prod_1’:
                    d=1
                    Z=np.linalg.eigh(A)[0]
                    Z[np.isnan(Z)]=0
                    Xhat = np.matmul((np.flipud(np.linalg.eigh(A)[1].T).T)[:
                                                             ,0:d],np.diag(np
                                                             .sqrt(np.flip(Z)
                                                             [0:d])))
                    Xhat[np.isnan(Xhat)] = 0
                    Matrix = np.matmul(Xhat,Xhat.T)
                elif met == ’dot_prod_5’:
                    d=5
                    Z=np.linalg.eigh(A)[0]
                    Z[np.isnan(Z)]=0
                    Xhat = np.matmul((np.flipud(np.linalg.eigh(A)[1].T).T)[:
                                                             ,0:d],np.diag(np
                                                             .sqrt(np.flip(Z)
                                                             [0:d])))
                    Xhat[np.isnan(Xhat)] = 0
                    Matrix = np.matmul(Xhat,Xhat.T)
                print(’Matrix_’+’n_’+str(n)+’_’+’r_’+rt+’_’+fun+’_’+met)
                np.savetxt(’matrices_simulation/’+’Matrix_’+’n_’+str(n)+’_’+
                                                         ’r_’+rt+’_’+fun+’_’+
                                                         met+’.csv’, Matrix,
                                                         delimiter=",")
```

Code for generating heatmap pictures:

```
def plot_triangular_combined_heatmaps(matrices, labels):
k = len(matrices)
vmin = np.min(matrices)
vmax = np.max(matrices)
assert k % 2 == 0, "Number of matrices (k) must be even."

# Number of combined heatmaps
num_heatmaps = k // 2

# Create a figure with subplots arranged horizontally
fig, axes = plt.subplots(1, num_heatmaps, figsize=(6 * num_heatmaps, 5))
```

```
    if num_heatmaps == 1:
        axes = [axes]  # Ensure axes is iterable for a single heatmap

    for i in range(num_heatmaps):
        # Extract the pair of matrices
        lower_matrix = matrices[2 * i]
        upper_matrix = matrices[2 * i + 1]
        n = lower_matrix.shape[0]  # Assuming square matrices

        # Create a combined matrix filled with zeros (or another placeholder
                                                 value)
        combined_matrix = np.zeros_like(lower_matrix)

        # Fill the lower triangle with the first matrix
        combined_matrix[np.tril_indices(n)] = lower_matrix[np.tril_indices(n)]

        # Fill the upper triangle with the second matrix
        combined_matrix[np.triu_indices(n)] = upper_matrix[np.triu_indices(n)]

        # Plot the combined matrix as a heatmap
        ax = axes[i]
        im = ax.imshow(combined_matrix, cmap='viridis', vmin=vmin, vmax=vmax,
                                                 interpolation='nearest')
        ax.set_title(f"{labels[2 * i]} (L) & {labels[2 * i + 1]} (U)", fontsize=
                                                 25)
        ax.axes.get_xaxis().set_ticks([])
        ax.axes.get_yaxis().set_ticks([])



    cbar = plt.colorbar(im, ax=ax)
    cbar.ax.tick_params(labelsize=25)
    plt.tight_layout()
    plt.show()
```

Example of code used to generate plots from simulations:

```
    #plot different methods at different sample sizes
NN = [100, 500]
methods = ['true', 'observed', 'HK1_c_1','HK1_c_2','HK2','HNN1', 'dot_prod_1', '
                                         dot_prod_5']
labels = ['true', 'observed', 'HK1','HK1_c2','HK2','HNN1', 'DP1', 'DP5']
rho_type = ['constant', 'high', 'boundary']
function = ['product', 'horseshoe', 'high density']
for n in NN:
    for rt in rho_type:
        for fun in function:
            matrices = []
            for met in methods:
                my_data = genfromtxt('Matrix_'+'n_'+str(n)+'_'+'r_'+rt+'_'+fun+'
                                                        _'+met+'.csv', delimiter
                                                        =',')
                matrices.append(my_data)
            plot_triangular_combined_heatmaps(matrices, labels, 'heatmap_'+'n_'+
                                                    str(n)+'_'+'r_'+rt+'_'+fun)
```

Chapter 2

Nonparametric bootstrap for exchangeable networks

Abstract

Inference on network data is challenging due to the strong dependence between observations, which renders standard techniques incorrect. To address this, we propose a valid bootstrap procedure for network data based on a nonparametric linking probabilities estimator. We prove that the distribution of the bootstrap network is consistent for the distribution of the original network in terms of a Wasserstein distance. We also provide conditions under which distributions of a class of functions related to U-statistics on the bootstrapped networks consistently replicate the distributions of the corresponding statistics on the original network. Monte Carlo simulations show good confidence interval coverage for a wider class of network functions than those accounted for by our theory. We apply our method to the data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013): we replicate their findings, but also show that our method works under weaker assumptions and with a significantly smaller sample size. Finally, we propose an alternative specification of their model which takes advantage of our linking probabilities estimator and may be of interest independently of our bootstrap procedure.

2.1 Introduction

Many papers in economics studying information diffusion (e.g. Banerjee, Chandrasekhar, Duflo, and Jackson (2013), Alatas et al. (2016)), impact of the most influential individuals (e.g. Banerjee, Chandrasekhar, Duflo, and Jackson (2019), Breza and Chandrasekhar (2019)), inherent features of networks (e.g. Chetty et al. (2022), Banerjee et al. (2024)) and other models on network data run into the issue that it is very difficult to conduct statistical inference on complex, interconnected data structures represented by networks. In this paper, we propose a solution: a bootstrap procedure which does not impose strong assumptions on the form of the network-generating function and can be applied to a wide range of network statistics.

The default approach when analysing the behaviour of statistics is finding an asymptotic approximation to their distribution, a technique easiest to apply to simple models and data with limited dependence. Unfortunately, the models built on networks are often complex and the networks themselves tend to exhibit a deeply interconnected structure. All individuals in a network are closely related: the concept of "six degrees of separation" shows that nearly all users of social media platforms like Facebook or Twitter are at most six connections away from each other, while the average distance is below four. At the same time, the number of connections grows more slowly than the network size. This phenomenon is known as sparsity and can be illustrated by the fact that, during their peak growth periods, social media platforms gained new users at a faster rate than individual users gained new connections. Because of the issues of strong connectedness, sparsity, and complicated functional forms of network statistics, asymptotic theory for network statistics tends to be complicated, specialised to certain classes of estimators, and, in many cases, still underdeveloped.

For similar reasons, standard bootstrap techniques are not valid for network data: there is a need for a specialised bootstrap procedure specifically designed to deal with this kind of dependence. The few existing methods for bootstrapping network data suffer from either limited applicability (they tend to focus on specific classes of network statistics and cannot be easily extended to e.g. regressions controlling for the dependence structure defined by the network) or restrictive parametric functional form for the components of the network-generating process. We address both of those concerns.

We propose a bootstrap procedure which takes a given network, uses it to approximate the data-generating process, and creates new networks with a similar structure to the original one. If we are interested in the distribution of a particular statistic of the original network, we can approximate it by estimating the same statistic on a large number of bootstrapped networks.

We assume a general form of a network-generating process in which the observed network is determined by an unknown distribution over types of individuals and an unknown function determining the the probability of a link between any pair of individuals. Under this assumption, each person is characterised by a set of (possibly unobserved) features that are independent of the features of others. The links are assumed to come from independent draws with probabilities determined by a binary linking function which takes the features of any two individuals as inputs and outputs the probability of a link between them.

If we knew the linking function, we could generate networks similar to the observed one by firstly resampling from the original set of individuals and then adding links based on probabilities determined by the linking function. However, as the linking function is unknown, we replace it with a consistent estimate which we proposed in Chapter 1. The estimator takes advantage of the information provided by the set of observed connections to identify similar individuals and to estimate linking probabilities based on a kernel-style estimator which takes a form of a local weighted average of the number of links between any person j and those similar (in terms of their linking behaviour) to person i.

Having developed a method to generate bootstrap networks, we can repeat any analysis we were doing on the original network on each of the bootstrap networks. This provides a bootstrap for network statistics. We provide conditions under which our procedure achieves consistency. We show this in two ways: we borrow a notion of Wasserstein distance between network generating distributions from Levin and Levina (2019) and we show that the distance between the bootstrap network generating process and the true network generating process goes to zero in probability as we increase the sample size. Unfortunately, this is not sufficient to ensure that the distribution of any statistic on a bootstrap network replicates the corresponding distribution of that statistic on the original network. We show this directly for a class of statistics which are closely related to U-statistics. The motivation for this is twofold: this is a wide class of functions and includes some estimators we may be directly interested in, for example the density of connections within a network. Additionally, this class includes motif densities, i.e. the densities of different patterns (e.g. triangles, stars, cycles of length m) on subgraphs of the adjacency matrix. These are sometimes referred to as “network moments” because they characterise the network generating distribution: if two networks match on densities of all possible patterns, they come from the same distribution. Hence proving that our bootstrap procedure correctly recovers the distributions of all motif densities implicitly shows that the bootstrap networks share the same asymptotic network generating distribution as the original network.

While we do not currently have explicit asymptotic theory for other classes of network functions, for example measures of centrality, clustering, eigenvalues of the adjacency matrix, or parameters of regressions on networks, our simulations suggest that our method is more widely applicable and can be used to recover distributions of these kinds of statistics.

In our application, we provide an illustration of how our bootstrap method can be extended to an information diffusion model over a network using data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013). Under the setup of the original paper, we are able to provide slightly narrower confidence intervals. Additionally, our method allows us to perform estimation on a significantly smaller sample: while the original paper relies on asymptotics in the number of networks and requires observing many villages, our method is asymptotic in the village size, meaning that we can construct confidence intervals given data on a single village. This has the potential to drastically lower data collection costs. We also propose an alternative model specification which uses our linking function estimator as a proxy for the strength of connection and which could be of interest independently of the bootstrap procedure.

In Section 2.2 we summarise the related literature. The setup of the model is described in Section 2.3, where we also provide our bootstrap procedure. Section 2.4 includes the statements of our main results: a Wasserstein distance convergence in Theorem 2.4.1 and a bootstrap consistency result for a specific class of estimators related to U-statistics in Theorem 2.4.2. Section 2.5 shows results of Monte Carlo simulations and Section 2.6 describes an application to the data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013). Section 2.7 concludes. The appendices start with a list of all notation. Section 2.A includes all proofs, Section 2.B provides the codes, additional tables, plots for simulations, Section 2.C includes extensions.

2.2 Related literature

2.2.1 Network Bootstrap

There are a few existing bootstrap procedures for different functions on networks. Most of the literature focuses on bootstrapping a class of network functions closely related to U-statistic, or their subset, motif densities (i.e. the proportions of subgraphs of a given size which take the form of a specific pattern or 'motif,' e.g. the proportion of subgraphs of size three which are fully connected, or the proportion of subgraphs of size four in which there is only one link). Our procedure can be applied to a much wider class of functions.

Green and Shalizi (2022) propose two types of bootstrap: the empirical bootstrap, in which they resample individuals and put a link between them if they were linked in the original graph, and a parametric histogram bootstrap. The empirical bootstrap can be seen as a special case of ours (with a very small bandwidth), it is simple and computationally attractive, but it suffers from a few types of bias: whenever an individual gets resampled more than once, these copies are not linked (as there were no self-links in the original graph), and they share the same link patterns with all other individuals (there is correlation between link formation in the bootstrap graph which was not present in the original graph). In our case, the resampled copies of the same individual form links with identical probabilities, but their link formation is independent, allowing their realised links to differ. We can also compute the probability of a link between two copies of the same individual, allowing them to be linked in the bootstrap graph. Green and Shalizi (2022) prove that this dependence and bias is asymptotically negligible for motif densities, but this is not necessarily true for other functions. Our simulations show that their procedure does not perform well e.g. for eigenvalues other than the highest one.

Levin and Levina (2019) assume a specific functional form of the linking function.[1] They propose two methods: one in which they directly estimate a U-statistic and one in which they generate a full network that can be used for estimating more general functions of a network, including eigenvalues and measures of small-world behaviour. This is the only paper we are aware of which provides results for functions of the entire network: they show that the entire bootstrapped network converges to an independent copy of the original network in terms of a new notion of Wasserstein network distance they define. Under our more general nonparametric specification we are able to show convergence in terms of the same distance (see Theorem 2.4.1).

Lin, Lunde, and Sarkar (2020) propose a computationally efficient multiplier bootstrap for motif densities, based on approximating the first (for large sparse graphs) or first and second (for smaller denser graphs) order terms of a Hoeffding decomposition of the U-statistic. Their method is specific to this class and, unlike ours, it cannot be extended to other types of network functions. They show higher-order accuracy of their quadratic bootstrap using an Edgeworth expansion. The theory of Edgeworth expansion for motif densities is developed Zhang and Xia (2022) who show higher order correctness of a studentised version of the empirical bootstrap of Green and Shalizi (2022). We believe similar methods could be used to show higher-order accuracy of our method, but we do not pursue this direction in the current work.

Shao and Le (2024) provide a parametric bootstrap in a setting different from all the previously mentioned papers, where the nodes are non-exchangeable. In our notation this corresponds to a situation in which the ξ and the matrix of link probabilities are fixed, $h_{0,n}$ takes a known parametric form, and the randomness comes only from η. Their analysis focuses on quantifying the bias and providing bias-corrected bootstrap procedure for motif densities.

The network setup is a special case of an exchangeable array.[2] Papers which propose bootstrap for exchangeable arrays include Davezies, D'Haultfœuille, and Guyonvarch (2021), whose

[1] They assume a random dot product graph with a linking function: $h_{0,n}(\xi_i, \xi_j) = \xi_i' \xi_j$ where ξ_i is a vector of latent positions which can be interpreted as characteristics of individual i.

[2] Using notation from Davezies, D'Haultfœuille, and Guyonvarch (2021), our model is a special case of an exchangeable and dissociated array with $k = 2$ and U_{ij} corresponding to the randomness due to Bernoulli trials $\tau(u_i, u_j, u_{ij}) = \mathbb{1}(h(u_i, u_j) \leq u_{ij})$. The kernels of U-statistics on networks can be represented as higher-dimensional ($2 \leq k < \infty$) exchangeable arrays.

method in our setting is identical to empirical bootstrap, and Menzel (2021), who proposes a new wild bootstrap procedure based on splitting the statistic of interest into orthogonal components, estimating them by sample analogues, and resampling each component with appropriate scaling. Menzel (2021) also points out that, depending on the dependence structure, the limiting distribution may be nonstandard. Their LLN and CLT results apply to functions of finite k-dimensional subgraphs which take a form of U- or V-statistics (including degenerate cases), their smooth functionals and Z-estimators. Their methods are local, they can be applied to functions of finite subgraphs and cannot account for dependence over the whole adjacency matrix, as in the case of eigenvalues or some centrality measures covered by our method.

In terms of the allowed level of sparsity, we impose a stronger requirement for acyclic motifs than both models in Green and Shalizi (2022) as well as Lin, Lunde, and Sarkar (2020), but our requirement is the same as for cycles in Lin, Lunde, and Sarkar (2020) and is weaker than that for general motifs for Green and Shalizi (2022) empirical graphon. In comparison with Green and Shalizi (2022) histogram graphon, our sparsity condition for general motifs becomes weaker only when $m > 4$, and we also impose weaker conditions than L-Lipschitz on the linking function. Levin and Levina (2019) only include sparsity considerations in one result, for acyclic motifs and cycles. Their assumption is weaker than ours, which is not surprising given their model is parametric.

Apart from bootstrap, other ways of estimating distributions of network statistics include the asymptotic theory for motif densities provided byBickel, Chen, and Levina (2011). Subsampling methods have been proposed by Bhattacharyya and Bickel (2015), who give results for motif densities, and Lunde and Sarkar (2022), who provide consistency results for general functions and specify them to two classes: motif densities and eigenvalues of graphons of finite rank. Their methods require minimal assumptions and allow for sparser graphs than ours.

2.2.2 Other

In the proofs that our bootstrap procedure is reliable we use a framework inspired by Politis et al. 1999. Our results can be seen as an extension of Bickel and Freedman (1981), the classic paper providing conditions for consistency of bootstrap for U-statistics. We extend their analysis to the case where the objective function becomes a U-statistic only after taking expectation conditional on a vector of unobserved characteristics and after substituting the true linking function for its estimator as the input to the kernel function. We show that our linking function estimator converges to the true linking function in a sense which is sufficient for the bootstrap equivalent to converge weakly in probability to the same limiting distribution as the object of interest in the original sample. Because of the additional levels of approximation, we achieve a weaker notion of convergence than convergence weakly almost surely (see Definition 2.4.3) in Bickel and Freedman (1981), but our result is still sufficient to provide asymptotically correct bootstrap confidence intervals.

For our empirical application, we use the data and some of the codes from Banerjee, Chandrasekhar, Duflo, and Jackson (2013). We confirm their results using our method and we repeat a part of their analysis under weaker assumptions: where the original paper performs the estimation aggregating over many villages, we are able to provide estimates and confidence intervals on individual village level. This way we relax the assumption that the parameter values are the same across all villages and we show that there is heterogeneity between them. We are also able to run a related model based on the strength of connections between household rather than the less informative binary information on presence or lack of connection. We find that removing this one level of approximation has a significant effect on our conclusions.

2.3 Model: setup, definitions and the bootstrap procedure

2.3.1 Setup

We follow the standard setup in the literature (see e.g. Green and Shalizi (2022), Bhattacharyya and Bickel (2015), Zeleneev (2020) and Auerbach (2022)) known as the latent space model. This is the same setup as in Chapter 1.

We observe an adjacency matrix A which corresponds to an undirected, unweighted graph on n nodes (also referred to as individuals) indexed by $i \in \{1, 2, \dots, n\}$. The matrix is symmetric, has zeros on the main diagonal and ones in positions corresponding to edges in the graph ($A_{ij} = 1$ if and only if there is an edge between nodes i and j). For a vector of index numbers $\iota = (\iota_1, \dots, \iota_m)'$ with $\iota_i \in \{1, 2, \dots, n\}$ we let $A(\iota)$ denote the corresponding submatrix defined on nodes in ι (i.e. A from which we remove rows and columns not in ι). Each node i is characterised by a vector of unobserved features[3] ξ_i, drawn independently from their common distribution F_0 with support $Supp(\xi_i)$. We denote the vector of all $\{\xi_i\}_{i=1}^n$ by ξ and we let $\xi(\iota) = (\xi_{\iota_1}, \dots, \xi_{\iota_m})$. We assume that the distribution has no point mass, i.e. for $\xi_i, \xi_j \sim F_0$ we have[4] $P_{F_0}(\xi_i = \xi_j) = 0$. We impose more assumptions[5] on F_0 in Assumption 1.2.

Let $h_{0,n} : Supp(\xi_i) \times Supp(\xi_i) \to [0, 1]$ be a symmetric, measurable linking function[6] which

[3] This corresponds to the vector of latent positions X_i in Levin and Levina (2019).

[4] This is without loss of generality: if we had a distribution with a point mass we could define a new support of ξ and a new F_0 in which the point mass would be replaced by a region of ξ of total measure equal to the probability at the original point.

[5] The assumptions are implicit and would be implied by F_0 bounded above and separated away from zero with $h_{0,n}$ piecewise Lipschitz.

[6] The linking function has been referred to as the coupling function $g(.,.)$ in Zeleneev (2020) and the graphon

can be decomposed as:

$$h_{0,n}(u,v) = \rho_n w_0(u,v) \tag{2.1}$$

where $\int w_0(u,v)dF_0(u)dF_0(v) = 1$.

For each pair of nodes i, j, $h_{0,n}(\xi_i, \xi_j)$ maps their unobserved characteristics ξ_i, ξ_j into the probability of a link (edge) between them, i.e. the probability with which $A_{ij} = 1$. For concise notation, we use $h_{0,n}(\xi(\iota))$ to mean the collection of pairwise linking probabilities between the elements of $\xi(\iota)$. We treat the linking function as unknown, making minimal assumptions on its properties in Assumption 1.2: we require that for each input there is a neighbourhood of sufficiently large measure in which the behaviour of the function remains similar. Importantly, we do not require a specific form (e.g. random dot product structure: $h_{0,n}(\xi_i, \xi_j) = \xi_i'\xi_j$ like in Levin and Levina (2019)), we do not impose any shape constraints (e.g. that the function is strictly increasing in its inputs).

The decomposition into ρ_n and w_0 can be seen as a normalisation which allows us to interpret ρ_n as the expected edge density (the marginal probability of an edge between two nodes). We assume $\rho_n \to 0$ as $n \to \infty$, which captures sparsity. We specify bounds on the rate at which ρ_n approaches zero which still allow us to reliably estimate parameters and their distributions. For the linking function estimator we require that the density decreases at a slower rate than $\sqrt{\frac{\log(n)}{n}}$ (see Assumption 1.1), while in other sections we may strengthen this requirement, e.g. in Theorem 2.4.2 the allowed level of sparsity depends on how complicated the statistic we are estimating is.

w_0 is the underlying linking/graphon function after accounting for sparsity. While w_0 cannot be interpreted directly as a probability, it has similar properties, e.g. it is bounded.[7] This is the function which determines the data generating process and the function the statistics of which we want to analyse. Although in a sample of size n we encounter its rescaled version $h_{0,n}$, for any asymptotic results we need to remove the effect of sparsity and we look at normalisations which are function of $\frac{h_{0,n}}{\rho_n}$.

To capture the way in which the linking function $h_{0,n}$ is translated into the observed links in A we introduce a random noise parameter: for $1 \leq i \leq j \leq n$ let $\eta_{ij} \overset{ind}{\sim} \mathcal{U}[0,1]$ be independent

function in Green and Shalizi (2022).

[7] This is a common assumption in the literature, though it is sometimes relaxed to allow $w_0(u,v) \in \mathbb{R}_+$ and let $h_{0,n}(u,v) = \min\{w_0(u,v), 1\}$. This affects the interpretation of ρ_n as the density and makes it more difficult to infer $h_{0,m}$ from $h_{0,n}$. Our results could be generalised to allow for unbounded w_0 at the expense of more complicated proofs and additional assumptions on bounded moments of w_0 or its functions.

of ξ. We denote the vector of η_{ij} by η. We assume:[8]

$$A_{ij} = A_{ji} = \mathbb{1}\left(h_{0,n}(\xi_i, \xi_j) \geq \eta_{ij}\right) \tag{2.2}$$

$$A_{ii} = 0. \tag{2.3}$$

Note that $E(A_{ij}|\xi_i, \xi_j) = P(A_{ij} = 1|\xi_i, \xi_j) = h_{0,n}(\xi_i, \xi_j) = \rho_n w_0(\xi_i, \xi_j)$. To distinguish between adjacency matrices based on the true and estimated/simulated inputs we sometimes explicitly write A as a function: $A(h_{0,n}(\xi), \eta)$.

2.3.2 The object of interest

We are interested in some property of the network and we have an estimate of this property which is a function of the adjacency matrix. In order to learn about the property of interest we want to approximate the distribution of its estimator. More precisely, let $f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)$, or $f_n(A)$ in short, be a function of the observed adjacency matrix A based on Bernoulli trials with probabilities determined by a linking function $h_{0,n}$ of i.i.d observations ξ from the distribution F_0, on the sparsity ρ_n, and on the distribution F_0 itself. The distribution we would like to approximate is:

$$J_n\left(t, h_{0,n}, F_0\right) = P\left(f_n\left(A\left(h_{0,n}\left(\xi\right), \eta\right), \rho_n, F_0\right) \leq t\right). \tag{2.4}$$

Example. *To fix ideas, suppose we want to learn about the density $\rho_n = E(A_{ij})$. We may want to test if it takes a specific value predicted by our theory, or we may wish to test if two networks (or perhaps the same network at two points in time) have the same density level. We can estimate the density using the density estimator from the observed adjacency matrix A:*

$$\hat{\rho}_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} A_{ij}. \tag{2.5}$$

We could use $f_n(A) = \hat{\rho}_n$ directly, or we could recentre and normalise the above expression:

$$f_n^{\rho_n}(A(h_{0,n}(\xi), \eta), \rho_n, F_0) = \frac{\sqrt{n}}{\binom{n}{2}\rho_n} \sum_{1 \leq i < j \leq n} A_{ij} - E_{F_0}(h_{0,n}(\xi_i, \xi_j)) \tag{2.6}$$

to get a function which has a well-defined asymptotic distribution. The results in Theorem 2.4.2 imply that $\hat{\rho}_n$ is consistent for ρ_n and $f_n^{\rho_n}(A(h_{0,n}(\xi), \eta), \rho_n, F_0)$ is asymptotically normal. The

[8] This is one specific way of achieving:

$$A_{ij}|\xi = A_{ji}|\xi \overset{ind}{\sim} \text{Bernoulli}\left(h_{0,n}(\xi_i, \xi_j)\right)$$
$$A_{ii} = 0$$

finite-sample distribution is non-trivial and depends on F_0.

Our goal is to find a good approximation to this finite-sample distribution, e.g. in order to form confidence intervals for ρ_n. We do it by defining estimators $\hat{h}_n$ of the linking function $h_{0,n}$; $\hat{F}_n$ of the distribution of ξ; and $\hat{\rho}_n$ of the density parameter. We use these estimates to form B bootstrap adjacency matrices $A\left(\hat{h}_n\left(\xi_b^*\right), \eta_b^*\right)$, where the bth bootstrap adjacency matrix is evaluated using $\hat{h}_n$ based on ξ_b^* from $\hat{F}_n$, the bootstrap equivalent of ξ, and η_b^* is the bootstrap equivalent of η.

We evaluate $f_n\left(A\left(\hat{h}_n\left(\xi^*\right), \eta^*\right), \hat{\rho}_n, \hat{F}_n\right)$ for B bootstrap samples to get the simulated distribution:

$$\hat{J}_{n,B}\left(t, \hat{h}_n, \hat{F}_n\right) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left(f_n\left(A\left(\hat{h}_n\left(\xi_b^*\right), \eta_b^*\right), \hat{\rho}_n, \hat{F}_n\right) \leq t\right). \tag{2.7}$$

For B large enough this provides an arbitrarily good approximation[9] to $J_n\left(t, \hat{h}_n, \hat{F}_n\right)$ and can be used to approximate $J_n\left(t, h_{0,n}, F_0\right)$.

In the reminder of this section we define all the estimators and the bootstrap procedure.

2.3.3 Linking probabilities estimator

To estimate linking probabilities we employ the estimator from Chapter 1. See there for details, here we just restate the definitions.

Our estimator takes the form of a kernel estimator (weighted average of the presence of links between individuals similar to i and j, weighted by the level of similarity to i). To capture the similarity we use the distance between i and j:

$$d_{ik} = \sqrt{E\left(E\left(\frac{A_{lm}}{\rho_n}\left(\frac{A_{im}}{\rho_n} - \frac{A_{km}}{\rho_n}\right)\middle|\xi_i, \xi_k, \xi_l\right)^2\middle|\xi_i, \xi_k\right)}.$$

We can estimate its normalised version from the observed adjacency matrix:

$$\rho_n^2 \hat{d}_{ik} = \sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(\frac{1}{n}\sum_{m=1}^{n} A_{lm}\left(A_{im} - A_{km}\right)\right)^2}.$$

We let $K(\cdot)$ be a kernel function and a_n be a bandwidth parameter. We can estimate $h_{0,n}(\xi_i, \xi_j)$ as:

[9]
$$P_{F_0}\left(\sup_t \left|\hat{J}_{n,B}(t, \hat{h}_n, \hat{F}_n) - J_n\left(t, \hat{h}_n, \hat{F}_n\right)\right| > \varepsilon\right) \leq 4\sqrt{2}e^{-2B\varepsilon^2},$$
see references on p.5 of Politis et al. (1999) for more details.

$$\hat{h}_n(\xi_i, \xi_j) = \frac{\tilde{h}_n(\xi_i, \xi_j) + \tilde{h}_n(\xi_j, \xi_i)}{2}$$

where

$$\tilde{h}_n(\xi_i, \xi_j) = \frac{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) A_{tj}}{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right)}.$$

We show that this estimator is uniformly consistent for $h_{0,n}$ in Theorem 1.

We propose a way to choose the bandwidth based on the observed sample: we pick a_n for which the kernel estimator best justifies the observed ones and zeros. We define a leave-one-out version of $\hat{h}_n$:

$$\tilde{h}_n^-(\xi_i, \xi_j) = \frac{\sum_{\substack{t=1 \\ t \neq i,j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) A_{tj}}{\sum_{\substack{t=1 \\ t \neq i,j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right)}$$

$$\hat{h}_n^-(\xi_i, \xi_j) = \frac{\tilde{h}_n^-(\xi_i, \xi_j) + \tilde{h}_n^-(\xi_j, \xi_i)}{2}.$$

and use it to obtain an estimate for the log-likelihood:

$$\ell(A, a_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \log\left(\hat{h}_n^-(\xi_i, \xi_j)\right) + (1 - A_{ij}) \log\left(1 - \hat{h}_n^-(\xi_i, \xi_j)\right).$$

We choose a_n which maximises the above expression to be our bandwidth:

$$\hat{a} = \max_{a_n} \ell(A, a_n).$$

2.3.4 Empirical distribution function estimator

The formation of matrix A is determined by an initial sample of ξ from F_0 and a linking probability between any pair of elements from ξ. We have defined a way to estimate the linking probabilities, but we still need a way to recreate the formation of ξ. This follows a very standard procedure, with one twist. Since the elements of ξ are i.i.d. from F_0, we should be able to use a standard bootstrap (resample from the values from the original sample, with replacement) to create a bootstrap equivalent. The non-standard part is that ξ_i are unobserved. We get around it by resampling not directly from the set of ξ_i, but from the set of original nodes: we let each bootstrap node correspond to one of the original nodes and we assign the set of characteristics of a bootstrap node (ξ_i^*) to be equal to the set of characteristics of the resampled original node.

The resulting distribution is an empirical distribution function $\hat{F}_n$ defined as the CDF which corresponds to the probability mass function:[10]

$$P(\xi_i^* = x) = \begin{cases} \frac{1}{n} & \text{if } x \in \{\xi_1, \ldots, \xi_n\} \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Each bootstrap node corresponds to one of the original nodes and inherits its characteristics. The result is the same as if we formed the set of bootstrap characteristics ξ^* by resampling from the original set of characteristics ξ, with replacement.

2.3.5 Nonparametric network bootstrap procedure

We now describe the bootstrap procedure for $f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)$. Just for this section, we introduce simplified notation for the matrix of estimated distances (D) and the matrix of estimated linking probabilities (H).

1. Calculate the distance between each pair of nodes $i, j \in \{1, 2, \ldots, n\}$:

$$D_{ij} = \frac{1}{n} \sum_{t=1}^{n} \left(\frac{1}{n} \sum_{s=1}^{n} A_{ts} \left(A_{is} - A_{js} \right) \right)^2 .$$

2. Calculate the optimal bandwidth parameter $\hat{a}$ as described in Eq. (1.13).

3. Calculate the probability of a link between each pair of nodes $i, j \in \{1, 2, \ldots, n\}$:

$$\hat{h}_n(\xi_i, \xi_j) = \frac{1}{2} \left(\frac{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{D_{it}}{\hat{a}}\right) A_{tj}}{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{D_{it}}{\hat{a}}\right)} + \frac{\sum_{\substack{t=1 \\ t \neq i}}^{n} K\left(\frac{D_{jt}}{\hat{a}}\right) A_{ti}}{\sum_{\substack{t=1 \\ t \neq i}}^{n} K\left(\frac{D_{jt}}{\hat{a}}\right)} \right)$$

4. Calculate the density estimate of the original graph: $\hat{\rho}_n$ as described in Eq. (2.5).

5. For each $b = 1, \ldots, B$:

   (a) draw an i.i.d. sample $\{\xi_{b,i}^*\}_{i=1}^n$ of size n from $\hat{F}_n$, i.e. resample from the original set of nodes $\{1, 2, \ldots, n\}$ with equal probabilities and with replacement, then assign the unobserved characteristics of the bootstrap node to be the same as the unobserved characteristics of its corresponding original node. Let $\xi_b^* = \left(\xi_{b,1}^*, \ldots, \xi_{b,n}^*\right)'$.

   (b) draw $\eta_{b,ij}^* \overset{ind}{\sim} \mathcal{U}[0,1]$ for $1 \leq i \leq j \leq n$.

[10] We would like to use the standard definition of an empirical distribution function:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(\xi_i < x\right)$$

but unfortunately ξ_i are not observed, and may in general not be scalar, hence this notation does not apply.

(c) form the bootstrap adjacency matrix A_b^*:

$$A_{b,ij}^* = A_{b,ji}^* = \mathbb{1}\left(\hat{h}_n(\xi_{b,i}^*, \xi_{b,j}^*) \geq \eta_{b,ij}^*\right) \tag{2.9}$$

$$A_{b,ii}^* = 0. \tag{2.10}$$

(e.g. if $\xi_{b,i}^* = \xi_t$, $\xi_{b,j}^* = \xi_s$ then $\hat{h}_n(\xi_{b,i}^*, \xi_{b,j}^*) = \hat{h}_n(\xi_t, \xi_s)$).

(d) calculate the object of interest on the bootstrap adjacency matrix:

$$f_n(A_b^*) \equiv f_n\left(A\left(\hat{h}_n(\xi_b^*), \eta_b^*\right), \hat{\rho}_n, \hat{F}_n\right). \tag{2.11}$$

6. Form a $(1-\alpha)\%$ confidence interval for $f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)$ by taking the interval between $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of $\{f_n(A_b^*)\}_{b=1}^B$.

For a description of how we used this procedure in simulations see Section 2.5. For the codes used in simulation see Section 2.B.1.

2.4 Main results

In this section we state our main results which characterise the conditions under which the distribution of the bootstrap network and our entire bootstrap procedure are consistent. Since we rely on the consistency of the liking probabilities estimator from Chapter 1, we operate under the same set of assumptions:

Assumption 1 (The Assumptions for Uniform Consistency of the Linking Function Estimator)**.** *We make the following assumptions:*

1.1 $\frac{1}{\rho_n} = o\left(\sqrt{\frac{n}{\log(n)}}\right)$.

1.2 Let $N(\xi_j, \delta) = \left\{\xi_k : \sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta\right\}$ *denote the neighbourhood of* ξ_j *of size* δ *and let* $\omega(\delta) = \inf_{\xi_j \in Supp(\xi_j)} P\left(\xi_k \in N(\xi_j, \delta) | \, \xi_j\right)$*. There exist some* $\alpha, C > 0$ *such that* $\omega(\delta) \geq \left(\frac{\delta}{C}\right)^{\frac{1}{\alpha}}$ *for all* $\delta > 0$*.*

1.3 $K(\cdot)$ *is a kernel function which is*

- *a continuous bounded probability density function (non-negative:* $K(u) \geq 0$*, integrates to 1:* $\int K(u)du = 1$*),*
- *non-zero on a bounded support: there exists a* $D \in \mathbb{R}$ *such that* $\forall |u| > D : K(u) = 0$*,*
- *positive close to 0: there exist positive constants* C_1, C_2 *such that* $K(u) \geq C_1$ *whenever* $|u| \leq C_2$*,*

- *Lipschitz continuous: there exists $C > 0$ such that $|K(u) - K(v)| \leq C|u - v|$.*

1.4 The bandwidth can be written as $a_n = \rho_n^4 b_n$ for some $b_n = o(1)$ and $\frac{1}{b_n} = o\left(\left(\frac{n\rho_n^2}{\log(n)}\right)^{\frac{\alpha}{1+2\alpha}}\right)$.

2.4.1 Consistency of the bootstrap procedure in terms of Wasserstein distance

To show that the distribution of the bootstrap network approaches that of the original network we follow the approach from Levin and Levina (2019) (see their Section 4 for more details and motivation). We start by defining an appropriate notion of convergence between network distributions. Firstly, let the graph matching distance be the proportion of edges that differ between two graphs after their vertices have been aligned to minimise the number of such differences:

Definition 2.4.1 (Graph matching distance)**.** *Let A_1, A_2 be two $n \times n$ adjacency matrices, Π_n be the set of $n \times n$ permutation matrices and let $\|A\|_{1,1} = \sum_{i=1}^n \sum_{j=1}^n |A_{ij}|$. The graph matching distance is:*

$$d_{GM}(A_1, A_2) = \min_{P \in \Pi_n} \binom{n}{2}^{-1} \frac{\|A_1 - PA_2P'\|_{1,1}}{2}. \tag{2.12}$$

Equipped with a distance between graphs we can define a distance between two distributions over graphs by using the Wasserstein distance:

Definition 2.4.2. *Let A_1, A_2 be the adjacency matrices of two random graphs on n vertices and let $\Gamma(A_1, A_2)$ be the set of all couplings of A_1 and A_2 (i.e. all joint distributions with marginal distributions matching those of A_1 and A_2). For $p \geq 1$ the Wasserstein p-distance is given by:*

$$W_p(A_1, A_2) = \inf_{\nu \in \Gamma(A_1, A_2)} \left(\int d_{GM}^p(A_1, A_2) d\nu\right)^{\frac{1}{p}}. \tag{2.13}$$

Theorem 2.4.1. *Let A be the observed adjacency matrix, let H be another adjacency matrix drawn independently from the distribution of A and let A^* be a bootstrap adjacency matrix derived from A. Under Assumption 1:*

$$W_p^p(A^*, H) = o_p(\rho_n). \tag{2.14}$$

The graph matching distance is an upper bound on the cut metric, which in turn metrises convergence of subgraph densities, hence Theorem 2.4.1 implies that all subgraph densities of

A^* converge to the same limit as those of H, proving that the bootstrap network distribution converges to the original network's distribution.

Remark. *As noted by Levin and Levina (2019), this notion of convergence is not sufficient to ensure that $f(A^*)$ converges to the same distribution as $f(H)$ for a general function $f(\cdot)$.*

2.4.2 Consistency of the bootstrap procedure for U-statistics

Because the result of Theorem 2.4.1 is not sufficient to guarantee that the distribution of a function of A^* is close to the distribution of the same function of A, we show this directly for an important class of functions with known limiting distributions.

Throughout this argument we use stars to denote the bootstrap equivalent, e.g. $\xi_i^* \sim \hat{F}_n$.

Appropriate notions of convergence

We start by introducing the definitions used in this section.

We choose f_n that has a distribution limit (usually a normal random variable), i.e. we assume $J_n(t, h_{0,n}, F_0) \Rightarrow J(t, w_0, F_0)$ for some non-degenerate distribution $J(t, w_0, F_0)$, where "$\Rightarrow$" denotes weak convergence. One convenient way to characterise weak convergence is though the following distance between measures: let P and Q be probability measures on a common metric space S equipped with a distance d_S and let

$$f(S) = \left\{ f : S \to \mathbb{R} : |f(x) - f(y)| \leq d_S(x, y), \sup_{x \in S} |f(x)| \leq 1 \right\}$$

be the set of (Lipschitz) continuous and bounded real-valued functions on S, then:

$$d_W(P, Q) \equiv \sup_{f \in f(S)} \left| \int f(x) dP(x) - \int f(x) dQ(x) \right|.$$

It can be shown[11] that $P_n \Rightarrow P$ if and only if $d_W(P_n, P) \to 0$ as $n \to \infty$.

In order to prove consistency of the bootstrap procedure we would like to show that the distribution $J_n\left(t, \hat{h}_n, \hat{F}_n\right)$, the bootstrap equivalent of $J_n(t, h_{0,n}, F_0)$, achieves the same asymptotic distribution $J(t, w_0, F_0)$. Unfortunately, the concept of weak convergence cannot be applied directly to the bootstrap statistic because both the bootstrap distribution $\hat{F}_n$ and the estimator $\hat{h}_n$ are random functions depending on the realisation of ξ, hence $J_n\left(t, \hat{h}_n, \hat{F}_n\right)$ and $d\left(J_n(t, \hat{h}_n, \hat{F}_n), J(t, w_0, F_0)\right)$ are also random. To proceed, we define two new concepts which generalise weak convergence to account for this randomness:

[11] See e.g. Proposition (M) in Chapter I of Hahn (1993)

Definition 2.4.3. *We say that* P_n converges weakly *to* P almost surely*, denoted by* $P_n \overset{a.s.}{\Rightarrow} P$*, if* $d_W(P_n, P) \xrightarrow{a.s.} 0$*. For* $X_n \sim P_n$*,* $X \sim P$ *we write* $X_n \xrightarrow{d} X$ almost surely*.*

Analogously,

Definition 2.4.4. *we say that* P_n converges weakly *to* P in probability*, denoted by* $P_n \overset{p}{\Rightarrow} P$*, if* $d_W(P_n, P) \xrightarrow{p} 0$*. For* $X_n \sim P_n$*,* $X \sim P$ *we write* $X_n \xrightarrow{d} X$ in probability*.*

Although this is a weaker requirement than regular convergence in distribution, Gine and Zinn (1990), who introduced the concept of weak convergence in probability, show that this notion is sufficient for the construction of asymptotically correct confidence intervals.

Distributions of intermediate terms

Since the elements of A exhibit dependence (e.g. A_{ij} and A_{jk} both depend on ξ_j), it is relatively difficult to work with $f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)$ directly. We can instead consider its expectation taken with respect to the Bernoulli trials with probabilities determined by $h_{0,n}(\xi)$:

$$\tilde{f}_n\left(h_{0,n}\left(\xi\right), \rho_n, F_0\right) \equiv E_{h_{0,n}}(f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)|\xi) \tag{2.15}$$

where we have taken the expectation over η and the remaining object becomes a function of the i.i.d. ξ_i.

Let $\tilde{J}_n$ denote the distribution of $\tilde{f}_n$:

$$\tilde{J}_n(t, h_{0,n}, F_0) = P_{F_0}\left(\tilde{f}_n(h_{0,n}(\xi), \rho_n, F_0) \leq t\right). \tag{2.16}$$

The limit of $\tilde{J}_n$ is easier to find than that of J_n, and the limits coincide if we can show that $f_n - \tilde{f}_n$ is negligible.

Remark. *We often work with conditional expectations and switch between variables that follow different distributions (e.g. the true distribution* F_0 *and the estimated empirical distribution* $\hat{F}_n$*). When we think it is beneficial to clarify, we add subscripts to the expectation operator indicating with respect to which distribution we are taking the expectation. For example,* $E_{h_{0,n}}(f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)|\xi) = \int f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0) d\eta$ *indicates that we are taking expectation with respect to the independent Bernoulli trials with probabilities determined by* $h_{0,n}$ *while* $E_{h_{0,n},F_0}(f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0))$ *denotes the expectation with respect to both the Bernoulli trials and the true distribution of* ξ*. The latter can also be written as* $E_{F_0}\left(\tilde{f}_n(h_{0,n}(\xi), \rho_n, F_0)\right)$*, where* $\tilde{f}_n(h_{0,n}(\xi), \rho_n, F_0)$ *has already been integrated over the Bernoulli trials, hence its randomness only comes from* F_0*, the true distribution of* ξ*.*

To illustrate the need for these intermediate terms we introduce an important class of statistics for which $\tilde{f}_n$ take the form of a U-statistic.

Definition 2.4.5. *Let ι be a set of m unique nodes, ξ be an n-dimensional vector of i.i.d. draws from a distribution F and η an n-dimensional vector of independent draws from $\mathcal{U}[0,1]$, and $\rho \in [0,1]$ be the sparsity level. Denote the adjacency matrix on the subgraph with nodes in ι by $A(h(\xi(\iota)), \eta(\iota))$.*

Let $g(A(\iota)) : \{0,1\}^{\binom{m}{2}} \to \mathbb{R}$ be a non-degenerate symmetric function from a subgraph on $m < \infty$ nodes to the real line such that $E_{h_{0,n},F_0}(g(A(\iota))) = \theta$, where θ is a parameter of interest. We can estimate θ on the whole network A by

$$\hat{\theta} = \frac{1}{\binom{n}{m}} \sum_{1 \le \iota_1 < \iota_2 < \cdots < \iota_m \le n} g(A(\iota)). \tag{2.17}$$

To get the corresponding $f_n(A(h(\xi), \eta), \rho, F)$ with a well-defined distribution we recentre and normalise the above expression:

$$f_n^U(A(h(\xi), \eta), \rho, F) = \frac{\sqrt{n}}{\binom{n}{m}\rho^{\tau(g)}} \sum_{1 \le \iota_1 < \cdots < \iota_m \le n} (g(A(h(\xi(\iota)), \eta(\iota))) - E_{h,F}(g(A(h(\xi(\iota)), \eta(\iota))))) \tag{2.18}$$

and

$$\tilde{f}_n^U(h(\xi), \rho, F) = \frac{\sqrt{n}}{\binom{n}{m}\rho^{\tau(g)}} \sum_{1 \le \iota_1 < \cdots < \iota_m \le n} (\tilde{g}(h(\iota)) - E_F(\tilde{g}(h(\iota)))) \tag{2.19}$$

where $\tilde{g}(h(\xi(\iota))) \equiv E_h(g(A(h(\xi(\iota)), \eta(\iota)))|\xi(\iota))$ and we choose $\tau(g)$ to get a normalisation for which there exists a non-degenerate bounded function $\tilde{\tilde{g}}$ such that

$$\frac{\tilde{g}(h_{0,n}(\iota))}{\rho_n^{\tau(g)}} = \tilde{\tilde{g}}(w_0(\iota)) + O(\rho_n).$$

The choice of $\tau(g)$ is quite simple: it is the smallest number of ones such that $g(\cdot)$ evaluated at a vector of $\tau(g)$ ones and $\binom{m}{2} - \tau(g)$ zeros is non-zero. The normalisation is not important for practical applications. We introduce it in the definition because it is necessary to get a well-defined asymptotic distribution (see Theorem 2.4.2, without the normalisation the limiting value of the $\hat{\theta}$ would be 0), but we do not need it if our interest is in constructing a confidence interval for θ. To see this, suppose θ and $\hat{\theta}$ are as above and let the bootstrap equivalent of the estimator be $\hat{\theta}_b^* = \frac{1}{\binom{n}{m}} \sum_{1 \le \iota_1 < \cdots < \iota_m \le n} g(A_b^*(\iota))$. We can calculate the estimator for B bootstrap adjacency matrices and find a confidence interval for θ as $\left[2\hat{\theta} - \hat{\theta}_U^*, 2\hat{\theta} - \hat{\theta}_L^*\right]$ where

$\hat{\theta}_L^* = q_{\frac{\alpha}{2}}\left(\left\{\hat{\theta}_b^*\right\}_{b=1}^B\right)$ and $\hat{\theta}_U^* = q_{1-\frac{\alpha}{2}}\left(\left\{\hat{\theta}_b^*\right\}_{b=1}^B\right)$. This way we get:

$$\begin{aligned}1-\alpha &\simeq P\left(2\hat{\theta}-\hat{\theta}_U^* < \theta < 2\hat{\theta}-\hat{\theta}_L^*\right)\\ &\simeq P\left(\frac{\sqrt{n}}{\hat{\rho}_n^{*\tau(g)}}\left(\hat{\theta}_L^*-\hat{\theta}\right) < \frac{\sqrt{n}}{\hat{\rho}_n^{\tau(g)}}\left(\hat{\theta}-\theta\right) < \frac{\sqrt{n}}{\hat{\rho}_n^{*\tau(g)}}\left(\hat{\theta}_U^*-\hat{\theta}\right)\right).\end{aligned}$$

The above confidence interval for θ is a close approximation to the confidence interval for $\frac{\sqrt{n}}{\hat{\rho}_n^{\tau(g)}}\left(\hat{\theta}-\theta\right)$ of the form $\left[\frac{\sqrt{n}}{\hat{\rho}_n^{*\tau(g)}}\left(\hat{\theta}_L^*-\hat{\theta}\right), \frac{\sqrt{n}}{\hat{\rho}_n^{*\tau(g)}}\left(\hat{\theta}_U^*-\hat{\theta}\right)\right]$. The consistency of $\hat{\rho}_n$ and $\hat{\rho}_n^*$ for ρ_n follows from the proof of Theorem 2.A.2.

Example. *We now show the relation between f_n and $\tilde{f}_n$ on an example. Suppose $m = 3$, e.g. $\iota = (1, 2, 3)$, and the function g only depends on two entries*[12] *in A: $g(A(\iota)) = g(A_{1,2}, A_{2,3})$.*

Conditional on ξ, the Bernoulli trials that determine the entries of A are independent. Hence, for example,

$$P(A_{ij} = 1, A_{jk} = 1|\xi) = P(A_{ij} = 1|\xi)P(A_{jk} = 1|\xi) = h_{0,n}(\xi_i, \xi_j)h_{0,n}(\xi_j, \xi_k).$$

It follows that

$$g(A_{1,2}, A_{2,3})|\xi = \begin{cases} g(0,0) & \textit{with probability } (1-h_{0,n}(\xi_1,\xi_2))(1-h_{0,n}(\xi_2,\xi_3)) \\ g(0,1) & \textit{with probability } (1-h_{0,n}(\xi_1,\xi_2))h_{0,n}(\xi_2,\xi_3) \\ g(1,0) & \textit{with probability } h_{0,n}(\xi_1,\xi_2)(1-h_{0,n}(\xi_2,\xi_3)) \\ g(1,1) & \textit{with probability } h_{0,n}(\xi_1,\xi_2)h_{0,n}(\xi_2,\xi_3). \end{cases}$$

The conditional expectation is a function of $h_{0,n}(\xi(\iota))$:

$$\begin{aligned}E(g(A_{1,2}, A_{2,3})|\xi) &\equiv \tilde{g}(h_{0,n}(\xi_1,\xi_2), h_{0,n}(\xi_2,\xi_3))\\ &= g(0,0)(1-h_{0,n}(\xi_1,\xi_2))(1-h_{0,n}(\xi_2,\xi_3))\\ &+ g(0,1)(1-h_{0,n}(\xi_1,\xi_2))h_{0,n}(\xi_2,\xi_3) + g(1,0)h_{0,n}(\xi_1,\xi_2)(1-h_{0,n}(\xi_2,\xi_3))\\ &+ g(1,1)h_{0,n}(\xi_1,\xi_2)h_{0,n}(\xi_2,\xi_3).\end{aligned}$$

If $g(0,0) \neq 0$, the first term on the right is $O(1)$ and dominates over the next terms. In this case we choose $\tau(g) = 0$. If $g(0,0) = 0$ but $g(0,1) \neq 0$ or $g(1,0) \neq 0$, the dominating

[12] For simplicity in this example we used a function which is not necessarily symmetric. Before plugging it into Eq. (2.18) we should symmetrise it in the following way:

$$\bar{g}(A(\iota)) = \frac{g(A_{1,2}, A_{2,3}) + g(A_{1,2}, A_{1,3}) + g(A_{1,3}, A_{2,3}) + g(A_{2,3}, A_{1,2}) + g(A_{1,3}, A_{1,2}) + g(A_{2,3}, A_{1,3})}{6}.$$

term is proportional to $(1 - h_{0,n}(\xi_i, \xi_j))h_{0,n}(\xi_j, \xi_k) = (1 - \rho_n w_0(\xi_i, \xi_j))\rho_n w_0(\xi_j, \xi_k) = O(\rho_n)$, hence we choose $\tau(g) = 1$ to normalise it. If $g(0,0) = g(0,1) = g(1,0) = 0$ but $g(1,1) \neq 0$, the dominating term is proportional to $h_{0,n}(\xi_i, \xi_j)h_{0,n}(\xi_j, \xi_k) = \rho_n^2 w_0(\xi_i, \xi_j)w_0(\xi_j, \xi_k) = O(\rho_n^2)$, hence the correct normalisation is $\tau(g) = 2$. When $\rho_n \to 0$, only the dominating term influences the limiting behaviour. We call this dominating term $\tilde{\tilde{g}}$. In this example:

$$\tilde{\tilde{g}}(h_{0,n}(\xi_1, \xi_2), h_{0,n}(\xi_2, \xi_3)) =$$

$$= \begin{cases} g(0,0) & \textit{if } g(0,0) \neq 0 \\ g(0,1)h_{0,n}(\xi_2, \xi_3) + g(1,0)h_{0,n}(\xi_1, \xi_2) & \textit{if } g(0,0) = 0, g(0,1) \neq 0, g(1,0) \neq 0 \\ g(1,1)h_{0,n}(\xi_1, \xi_2)h_{0,n}(\xi_2, \xi_3) & \textit{if } g(0,0) = g(0,1) = g(1,0) = 0, g(1,1) \neq 0. \end{cases}$$

In all cases we have:

$$\frac{\tilde{g}(h_{0,n}(\xi_1, \xi_2), h_{0,n}(\xi_2, \xi_3))}{\rho_n^{\tau(g)}} = \tilde{\tilde{g}}(w_0(\xi_1, \xi_2), w_0(\xi_2, \xi_3)) + O(\rho_n)$$

where the first term is $O(1)$ and does not depend on the sample size n.

We will later do Taylor expansion of $\tilde{g}$, for which it is interesting to note that regardless of whether g is a nice function (continuous, differentiable, etc.) of A, $\tilde{g}$ is (infinitely many times) continuously differentiable in $h_{0,n}$ and has bounded derivatives. Taking a derivative of $\tilde{g}$ with respect to $h_{0,n}$ lowers the power on the $h_{0,n}$ terms by one, hence if $\tilde{g} \sim \rho_n^{\tau(g)}$ then $\tilde{g}' \sim \rho_n^{\tau(g)-1}$ and $\frac{\tilde{g}'}{\rho_n^{\tau(g)}} = \frac{1}{\rho_n}$.

Then

$$\tilde{f}_n(h_{0,n}(\xi), \rho_n, F_0) = \frac{\sqrt{n}}{\binom{n}{m}\rho_n^{\tau(g)}} \sum_{1 \leq \iota_1 < \iota_2 < \cdots < \iota_m \leq n} \left(\tilde{\tilde{g}}(h_{0,n}(\iota)) - E_{F_0}(\tilde{\tilde{g}}(h_{0,n}(\iota)))\right). \qquad (2.20)$$

We can think of the above as a function of the i.i.d. ξ. If g is symmetric, so is $\tilde{g}(h_{0,n}(\cdot))$, and the $\tilde{f}_n(h_{0,n}(\xi), \rho_n, F_0)$ takes the form of a (normalised) U-statistic, for which we have results such as LLN and CLT. Hence it is much easier to work with than the original $f_n(A(h_{0,n}(\xi), \eta), \rho_n, F_0)$ (which was not a U-statistic due to the dependence in A).

Remark. It is usually not the case that $\tilde{f}_n$ has the same form as f_n with A_{ij} replaced with $h_{0,n}(\xi_i, \xi_j)$, but it can happen in some special cases. One such example are motif densities, for which the g function is a product of terms of the form A_{ij} (if the motif has an edge between nodes i and j) and $(1 - A_{ij})$ (if the edge is supposed to be missing). For example, if the motif of interest is a triangle, we have $g(A_{ij}, A_{jk}, A_{ki}) = A_{ij}A_{jk}A_{ki}$. This is 1 if all inputs are equal to 1 and 0 in all other cases, hence $\tilde{g}(h_{0,n}(\xi_i, \xi_j), h_{0,n}(\xi_j, \xi_k), h_{0,n}(\xi_k, \xi_i)) = h_{0,n}(\xi_i, \xi_j)h_{0,n}(\xi_j, \xi_k)h_{0,n}(\xi_k, \xi_i)$.

Remark. *Because in the proof of Theorem 2.4.2 we rely on a CLT for U-statistics applied to $\tilde{f}_n$ instead of f_n, the normalisation by $\frac{1}{\rho_n^{\tau(g)}}$ is chosen to balance the rate of growth of the variance of $\tilde{g}$ rather than g. Take a simple example of $g(A_{ij}) = A_{ij}$. Then $\tau(g) = 1$ and:*

$$E\left(\left(\frac{g(A_{ij})}{\rho_n^{\tau(g)}}\right)^2\right) = \frac{E\left(h_{0,n}(\xi_i, \xi_j)\right)}{\rho_n^{2\tau(g)}} = \frac{1}{\rho_n^{\tau(g)}} \to \infty$$

$$E\left(\left(\frac{E\left(g(A_{ij})|\,\xi\right)}{\rho_n^{\tau(g)}}\right)^2\right) = \frac{\left(E\left(h_{0,n}(\xi_i, \xi_j)\right)\right)^2}{\rho_n^{2\tau(g)}} = O(1).$$

The following notation simplifies the statement of the theorem:

Definition 2.4.6. *Set $\mathcal{M}_m$: Let $\mathcal{M}_m$ be the set of all possible multisets*[13] *of cardinality m with elements from $\{1, 2, \ldots, m\}$.*

That is, $\mathcal{M}_m$ contains all possible combinations of index numbers from 1 to m that are of length m and can be all unique or have any value repeated any number of times.

We are now ready to state the final main result, which shows that for statistics which can be represented as in Definition 2.4.5 and are non-degenerate (i.e. $\sigma_1^2 \neq 0$): 1. the limiting distribution in probability of the bootstrap statistic is asymptotically normal and the same as the limiting distribution of the original statistic and 2. the bootstrap is consistent, in the sense that the finite-sample distribution of the bootstrap statistic approaches the finite-sample distribution of the original statistic as the sample size increases.

Theorem 2.4.2. *Let $f_n^U(A(h(\xi), \eta), \rho, F)$ be as in Eq. (2.18). There exists a normalisation*[14] *$\tau(g)$ and a function $\tilde{\tilde{g}} : Supp(\xi)^m \to \mathbb{R}$ such that $\frac{\tilde{g}(h_{0,n}(\xi(\iota)))}{\rho_n^{\tau(g)}} = \tilde{\tilde{g}}(w_0(\xi(\iota))) + O\left(\rho_n\right)$ and $0 < E\left(|\tilde{\tilde{g}}(w_0(\xi(j)))|\right) < \infty$ for all $j \in \mathcal{M}_m$. If Assumption 1 holds and:*

$$Var_{F_0}(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi(\iota)))|\xi_{\iota_1}\right)) \equiv \sigma_1^2 > 0$$

$$\frac{n}{\binom{n}{m}\rho_n^{\tau(g)}} \to 0$$

then

1. $f_n^U(A(h_{0,n}(\xi), \eta), \rho_n, F_0) \xrightarrow{d} N(0, m^2\sigma_1^2)$ *and*
   $f_n^U\left(A\left(\hat{h}_n\left(\xi^*\right), \eta^*\right), \hat{\rho}_n, \hat{F}_n\right) \xrightarrow{d} N(0, m^2\sigma_1^2)$ *in probability.*

2. $\sup_t \left|P\left(f_n^U\left(A\left(\hat{h}_n\left(\xi^*\right), \eta^*\right), \hat{\rho}_n, \hat{F}_n\right) \leq t\right) - P\left(f_n^U\left(A\left(h_{0,n}\left(\xi\right), \eta\right), \rho_n, F_0\right) \leq t\right)\right| \xrightarrow{p} 0.$

[13] A multiset is like a set but allows for repeated elements.

[14] For $m = 2$, if $g(0) \neq 0$ we set ${\rho_n}^{-\tau(g)} = 1$, $\tilde{\tilde{g}}\left(w_0\left(\xi_i, \xi_j\right)\right) = g(0)$ and if $g(0) = 0$ but $g(1) \neq 0$ we set ${\rho_n}^{-\tau(g)} = \frac{1}{\rho_n}$ and $\tilde{\tilde{g}}\left(w_0\left(\xi_i, \xi_j\right)\right) = g(1)w_0\left(\xi_i, \xi_j\right)$. More generally, for $m \geq 2$, ${\rho_n}^{-\tau(g)} = \frac{1}{\rho_n^k}$ where k is the smallest number of ones such that $g(\cdot)$ evaluated at a vector of k ones and $\binom{m}{2} - k$ zeros is non-zero.

In the above theorem we add two new assumptions on top of those in Assumption 1. The first one, $\sigma_1^2 \neq 0$, restricts our attention to non-degenerate U-statistics. Levin and Levina (2019) claim that in the case of degenerate U-statistics the approximation error is of a comparable size to the leading term, implying that their bootstrap cannot recover the distributions of degenerate U-statistics. As explained in Serfling (2009) section 5.5, in the degenerate case the correct normalisation would be of the form $\frac{n^{\frac{c}{2}}}{\binom{n}{m}\rho_n^2}$ for some $c \geq 2$ and we would expect a more complicated limiting distribution than normal. In that case $f_n - \tilde{f}_n = O\left(\sqrt{\frac{n^c}{\binom{n}{m}\rho_n^2}}\right)$, which could go to zero sufficiently fast to remain negligible. We suspect that recovering distributions of degenerate U-statistics could still be possible with our method but we leave the detailed analysis for future work.

The other condition: $\frac{n}{\binom{n}{m}\rho_n^{\tau(g)}} \to 0$ gives a restriction on the allowed level of sparsity. We require $\frac{1}{\rho_n} = o\left(n^{\frac{m-1}{\tau(g)}}\right)$, where $\tau(g) \in \{0, 1, \ldots, \binom{m}{2}\}$. For sufficiently large $\tau(g)$ this condition may be stronger than Assumption 1.1. This is not surprising: large $\tau(g)$ means that the function g takes non-zero values only for very rare events, and these events are even less common in sparser graphs. Hence to be able to maintain consistency we need to restrict the allowed level of sparsity.

This condition is needed to ensure that the $f_n - \tilde{f}_n$ term does not affect the limiting distribution. If $\frac{n}{\binom{n}{m}\rho_n^{\tau(g)}} = O(1)$, the limit of this term would affect the resulting distribution and the overall limit would be the current one plus the limit of this adjustment term. If $\frac{n}{\binom{n}{m}\rho_n^{\tau(g)}} \to \infty$ this adjustment term would dominate the asymptotic behaviour. In that case we would need to use normalisation by $\frac{n}{\binom{n}{m}\rho_n^{\frac{\tau(g)}{2}}}$. The currently dominating term under the new normalisation would go to zero. Deriving the distribution of the new dominating term is difficult due to a high level of dependence between the elements of A.

Remark. *Green and Shalizi (2022) specify the maximal allowed level of sparsity in two cases: when the motif is acyclic they assume $\frac{1}{\rho_n} = o(n)$ and for a general motif they require $\frac{1}{\rho_n} = o\left(n^{\frac{1}{2m}}\right)$ for the empirical graphon and the weaker condition of $\frac{1}{\rho_n} = o\left(n^{\frac{2}{m}}\right)$ for a general linking function estimator, e.g. their histogram graphon.*

These conditions are weakly stronger than our $\frac{1}{\rho_n} = o\left(n^{\frac{m-1}{\tau(g)}}\right)$: when $g(\cdot)$ corresponds to an acyclic motif[15] we have $\tau(g) \leq m-1$, hence $\frac{m-1}{\tau(g)} \geq 1$; when $g(\cdot)$ corresponds to a general motif[16] we have $\tau(g) \leq \binom{m}{2}$, hence $\frac{m-1}{\tau(g)} \geq \frac{2}{m} \geq \frac{1}{2m}$.

However, for consistency of our linking function estimator we require $\frac{1}{\rho_n} = o\left(\sqrt{\frac{n}{\log(n)}}\right)$, which is stronger than the $\frac{1}{\rho_n} = o(n)$ condition for acyclic motifs. For general motifs, our condition is always weaker than the condition needed for the empirical graphon. In comparison

[15] $\tau(g)$ corresponds to the number of edges in the motif and m denotes the number of vertices. The maximal number of edges in an undirected acyclic graph on m nodes is $m-1$.

[16] The maximal number of edges in an undirected graph on m nodes is $\binom{m}{2}$.

with $\frac{1}{\rho_n} = o\left(n^{\frac{2}{m}}\right)$ for histogram graphon, our condition is weaker when $m > 4$ and stronger for $m \leq 4$.

One of the motivations for looking at this class of functions on networks is that it contains subgraph densities, which can be viewed as 'network moments,' in the sense that if two networks match on the densities of all subgraphs they come from the same network generating distribution. Theorem 2.4.2 implicitly shows that the bootstrap network distribution converges to the distribution of the original network. However, as the subgraphs become more complicated we need to impose stronger conditions on sparsity, meaning that full convergence of all subgraphs would only follow for dense models in which ρ_n does not go to 0.

There are other linking function estimators (e.g. Zhang, Levina, and Zhu (2017)) and alternative ways to resample nodes. The next result characterises the conditions needed for consistency of the class of functions considered in Theorem 2.4.2 when we replace the $(\hat{h}_n, \hat{F})$ in our procedure with alternative estimators of $(h_{0,n}, F_0)$.

Lemma 2.4.1. *Theorem 2.4.2 holds for any estimators (h_n, F_n) of $(h_{0,n}, F_0)$ which satisfy:*

1. $E_{F_n}\left(\left(\frac{1}{\rho_n}\left(h_n(\xi_i^*, \xi_j^*) - h_{0,n}(\xi_i^*, \xi_j^*)\right)\right)^2\right) \xrightarrow{p} 0.$
2. $E_{F_n}\left(f\left(\xi^*(\iota)\right)\right) \xrightarrow{p} E_{F_0}\left(f\left(\xi(\iota)\right)\right)$ *for all* $f : Supp(\xi)^k \to \mathbb{R}$ *such that* $E_{F_0}\left(\left|f\left(\xi(\iota)\right)\right|\right) < \infty$ *for all* $\iota \in \mathcal{M}_k$*, for any* $k \leq 2m-1$.

In the proofs in Section 2.A we restate Theorem 2.4.2 in a generalised way which incorporates the conditions given in Lemma 2.4.1.

A consequence of Theorem 2.4.2 is that the bootstrap procedure can consistently recover critical values and asymptotically valid confidence intervals, as stated in Corollary 2.4.1. In order to be able to define the confidence intervals and comment on their coverage we need an inverse of the bootstrap distribution, but $J_n\left(t, \hat{h}_n, \hat{F}_n\right)$ may not necessarily be continuous or strictly increasing in t, hence it may not be invertible in the standard sense. Because of this we define the inverse of a distribution in the following way:

Definition 2.4.7. *Let:*

$$J^{-1}(\alpha, h, F) \equiv \inf\left\{t : J(t, h, F) \geq \alpha\right\} \tag{2.21}$$

be the αth quantile of the distribution $J(t, h, F)$.

Corollary 2.4.1. *Under the conditions of Theorem 2.4.2:*

1. $J_n^{-1}\left(1-\alpha, \hat{h}_n, \hat{F}_n\right) \xrightarrow{p} c_{1-\alpha}$*, where $c_{1-\alpha}$ is the $1-\alpha$ critical value*[17] *from $N(0, m^2\sigma_1^2)$.*

[17] I.e. $\Phi\left(\frac{c_{1-\alpha}}{m^2\sigma_1^2}\right) = 1-\alpha$ where $\Phi(\cdot)$ denotes the CDF of $N(0,1)$.

2. *If F_0 does not enter the function f_n^U directly but only through a parameter*[18] θ*:*

$$f_n^U(A(h_{0,n}(\xi),\eta),\rho_n,\theta),$$

then the $(1-\alpha)$ *confidence interval for* θ *constructed as:*

$$CI_n\left(1-\alpha, A, \hat{h}_n, \hat{F}_n\right) = \left\{\theta : J_n^{-1}\left(\frac{\alpha}{2}, \hat{h}_n, \hat{F}_n\right) \leq f_n^U(A, \hat{\rho}_n, \theta) \leq J_n^{-1}\left(1-\frac{\alpha}{2}, \hat{h}_n, \hat{F}_n\right)\right\} \tag{2.22}$$

is asymptotically valid:

$$P_{h_{0,n},F_0}\left(\theta \in CI_n\left(1-\alpha, A, \hat{h}_n, \hat{F}_n\right)\right) \xrightarrow{p} 1-\alpha. \tag{2.23}$$

In defining the bootstrap statistic and forming confidence intervals we do not normalise the $f_n(\cdot)$ by the estimated variance. This is in part because the variance estimators may not be readily available, and even when they are, they tend to be complicated (e.g. for the subclass of motif densities Green and Shalizi (2022) Lemma 2 gives an expression for an estimator of variance. It is derived combinatorially by considering all motifs that can be achieved by merging two copies of the motif of interest on partially overlapping sets of nodes). Another reason is that, as pointed out by Hahn (1993), convergence weakly in probability ensures convergence of moments over the set of bounded and Lipschitz continuous functions, which does not include $f(x) = x^2$, meaning that weak convergence in probability of our bootstrap estimator does not guarantee the consistency of its variance. When properties of the variance estimate are unknown, it is safer to use the percentile method for the construction of confidence intervals.

However, when a reliable variance estimate is known, normalising the statistic of interest by the estimate of its standard deviation could improve the performance of the bootstrap procedure. While we do not analyse the rates theoretically, the logic should be close to the case of standard bootstrap, where Edgeworth expansion arguments show that a normalised bootstrap with a pivotal limiting distribution can achieve a faster rate of convergence, see e.g. Hansen (2014) sections 10.8-10.11.

2.5 Simulations

We test the performance of our procedure using Monte Carlo simulations. We simulate the true adjacency matrices for $\xi_i \overset{iid}{\sim} \mathcal{U}[0,1]$ and one of the following linking functions:

1. dot product function: $h(\xi_i, \xi_j) = \rho_n \xi_i \xi_j$. This is the parametric form assumed by Levin

[18] For example in equation (2.18) we have $\theta = E_{h_{0,n},F_0}(g(A(\iota)))$.

and Levina (2019), it is a relatively simple function and a good benchmark.

2. horseshoe function: $h(\xi_i, \xi_j) = \frac{\rho_n}{2}\left(e^{-200\left(\xi_i - \xi_j^2\right)^2} + e^{-200\left(\xi_j - \xi_i^2\right)^2}\right)$. This function was also used by Green and Shalizi (2022). They borrow it from Wang (2016), who described it as "a challenging example for graphon estimation."

3. high-density function:
$h(\xi_i, \xi_j) = \frac{\rho_n}{0.975}\left(1 - \mathbb{1}\left(\left|\frac{1}{2} - \xi_i\right| \leq \frac{1}{20}\right)\mathbb{1}\left(\left|\frac{1}{2} - \xi_j\right| \leq \frac{1}{20}\right)\right)\left(1 - \frac{1}{2}\left(\left|\frac{1}{2} - \xi_i\right| + \left|\frac{1}{2} - \xi_j\right|\right)\right)$.
The previous two functions had relatively low density (by construction, $\rho_n \leq 0.25$ for the dot product function and $\rho_n \leq 0.113$ for the horseshoe function). This final function has $\rho_n \leq 0.759$, allowing us to test the performance with higher density levels.

The plots of these functions are included in Section 2.B.3.

In the estimation procedure we use the normal kernel: $K(u) = e^{-\frac{u^2}{2}}$ and the bandwidth $\hat{a}$ chosen by maximising $\ell(A, a_n)$, as described in Section 1.3.4.

We test the performance of the algorithm for a range of statistics which have economic interpretation, some of them are covered by our Theorem 2.4.2 while other are not, including some which are much more complicated to compute.

- density: $f_n(A) = \frac{1}{\binom{n}{2}} \sum\sum_{1 \leq i < j \leq n} A_{ij}$, i.e. the number of edges divided by the number of possible edges. This function is useful for normalisation and is an example of a U-statistic.

- triangle density: $f_n(A) = \binom{n}{3}^{-1} \sum\sum\sum_{1 \leq i < j < k \leq n} A_{ij}A_{jk}A_{ik}$, i.e. the proportion of all subsets of 3 nodes that are fully connected. This is another example of a U-statistic, but of a more complicated form. It can be used to measure clustering.

- transitivity: $f_n(A) = 3\frac{\#triangles}{\#triads} = \frac{tr\left(A^3\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}(A^2)_{ij} - tr(A^2)}$, i.e. the ratio of fully connected triples to connected triples. This statistic can be seen as the extent of triadic closure (the tendency of people who have a common friend to become friends with each other) and can be used as a measure of the ability of a group of people to maintain cooperation.

- kth largest eigenvalue of the adjacency matrix: $f_n(A) = \lambda_k(A)$. The eigenvector of the largest eigenvalue can be used as a measure of centrality, or the level of influence of individuals in a network. The other eigenvalues are also informative, e.g. the second eigenvalue of a normalised adjacency matrix, known as spectral homophily, can be seen as a measure of cohesiveness (Chetty et al. 2022), similar to transitivity.

- maximal betweenness centrality:
$f_n(A) = \max_i \sum_{j,k} \frac{\#\ shortest\ paths\ between\ j\ and\ k\ through\ i}{\#\ shortest\ paths\ between\ j\ and\ k}$. This is another measure of

how influential a person is, how well they are connected, which is an important determinant of, for example, how effective they would be at spreading information (Banerjee, Chandrasekhar, Duflo, and Jackson 2019) or holding someone accountable in their saving goal (Breza and Chandrasekhar 2019).

- modularity of the Louvain community detection algorithm: for a given partition of nodes into communities, modularity is defined as the proportion of edges within communities minus the proportion if the edges were distributed at random. The Louvain community detection algorithm aims to find a partition which maximises modularity by iteratively moving nodes to communities and aggregating communities until no further improvement is possible. We used functions 'louvain_communities' and 'modularity' from the Python networkx package, see their documentation for precise definitions.

A difficulty in running Monte Carlo simulations is that for each confidence interval coverage we need to generate 1000 true graphs, and for each of those we need 1000 bootstrap graphs, hence for each data point in our plots and tables we need to evaluate the statistic of interest a million times. Some of the above statistics take a bit of time to estimate, making the simulation process slow. To overcome this issue and to speed up the simulations we obtain the confidence interval coverage using the WARP procedure from Giacomini, Politis, and White (2013). The idea is that the distribution of *deviations* between the true graph and its bootstrap version should be similar for each true graph, hence we can generate 1000 true graphs, get only one bootstrap graph for each of them, and calculate the deviations for each true-bootstrap pair. Then we can pretend we got 1000 bootstrap graphs for each true graph by adding these deviations to the true graph. It is a clever trick which allows us to generate only one bootstrap graph for each true graph, making the computation time close to 500 times faster. The procedure is as follows:

1. Generate S true adjacency matrices on n nodes using the same true linking function (usually $S = 1000$, n between 25 and 1000).

2. For each true adjacency matrix A_s:

   (a) Find the optimal bandwidth $\hat{a}_s = \max_a \ell(A_s, a)$.

   (b) Calculate the matrix $\hat{h}_{n,s}$ based on A_s with bandwidth $\hat{a}_s$.

   (c) Resample n nodes of A_s to form the nodes of the bootstrap graph.

   (d) Generate a single bootstrap adjacency matrix $A^*_{s,1}$ by adding an edges between nodes with probabilities determined by $\hat{h}_{n,s}$.

   Note that the number of bootstrap replications is $B = 1$.

3. Estimate the true value by the average of the statistic evaluated for the true graphs: $f_n^{(true)} = \frac{1}{S}\sum_{s=1}^{S} f_n(A_s)$ (or use the theoretical true value, if known).

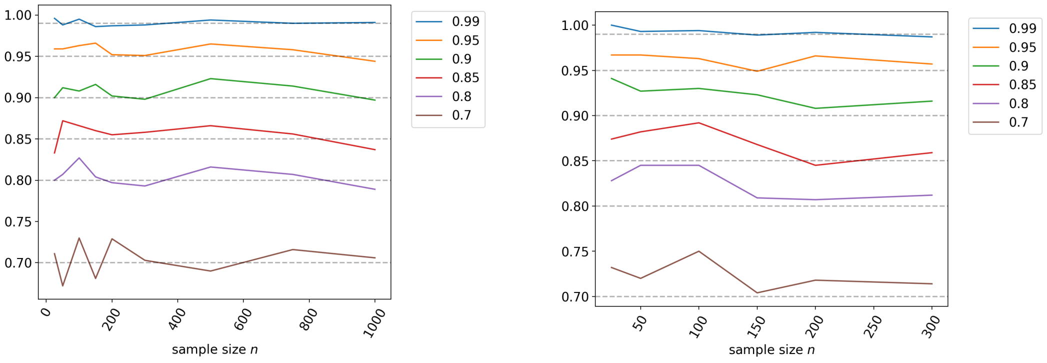
4. Calculate the deviation of the statistic in the bootstrapped graph from the statistic evaluated for the corresponding true graph: $f_n(A^*_{s,1}) - f_n(A_s)$ for all $s \in \{1,\dots,S\}$. Denote the αth quantile of the empirical distribution of this set by $\hat{q}_\alpha\left(\left\{f_n(A^*_{s,1}) - f_n(A_s)\right\}_{i=1}^{S}\right)$.

5. Calculate the confidence intervals as: $CL_s = [CL_s^l, CL_s^u]$, where

$$
\begin{aligned}
CL_s^l &= f_n(A_s) - \hat{q}_{1-\frac{\alpha}{2}}\left(\left\{f_n(A^*_{s,1}) - f_n(A_s)\right\}_{i=1}^{S}\right)\\
CL_s^u &= f_n(A_s) - \hat{q}_{\frac{\alpha}{2}}\left(\left\{f_n(A^*_{s,1}) - f_n(A_s)\right\}_{i=1}^{S}\right).
\end{aligned}
$$

6. Store the empirical coverage, i.e. the proportion of confidence intervals which cover the true value: $1-\alpha^{(emp)} = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\left(CL_s^l \leq f_n^{(true)} \leq CL_s^u\right)$.

The following plots and tables show results of some of our simulations. The code used in the simulations can be found in Section 2.B.1 and tables with more results are in Section 2.B.2.



(a) Confidence interval coverage for density.

(b) Confidence interval coverage for λ_1.

Figure 2.1: Confidence interval coverage for different sample sizes n based on Monte Carlo simulations using the product generating function and $\rho_n = 0.1875$.

We start by looking at the confidence interval coverage for different values of n, ρ_n and α. From Fig. 2.1 we can see that performance at different α is quite similar, but larger values have proportionally larger deviations and allow us to see the trend more clearly. This is why, although in practice we tend to be most interested in the 95% confidence intervals, we present results for 70% or 80% in most of our plots. From Fig. 2.2 we can see that at a constant density level (no sparsity, ρ_n does not decrease with n) the performance improves with sample size. For $n \geq 250$ all statistics achieve good confidence interval coverage levels, although not always perfect: we may get coverage of e.g. 60% instead of 70%. As the sparsity level increases ($\rho_n \to 0$), the performance tends to get worse, but remains close to desired for statistics which are easier to estimate (e.g. density, triangle density, or the highest eigenvalue) while it gets significantly worse for more complicated statistics (e.g. λ_{10}). To some extent, we may be able to overcome these issues by changing the choice of bandwidth, as we describe below. The bottom two panels check performance for sparsity levels at which our theoretical results do not give any performance guarantees: as expected, the performance is poor in those cases.

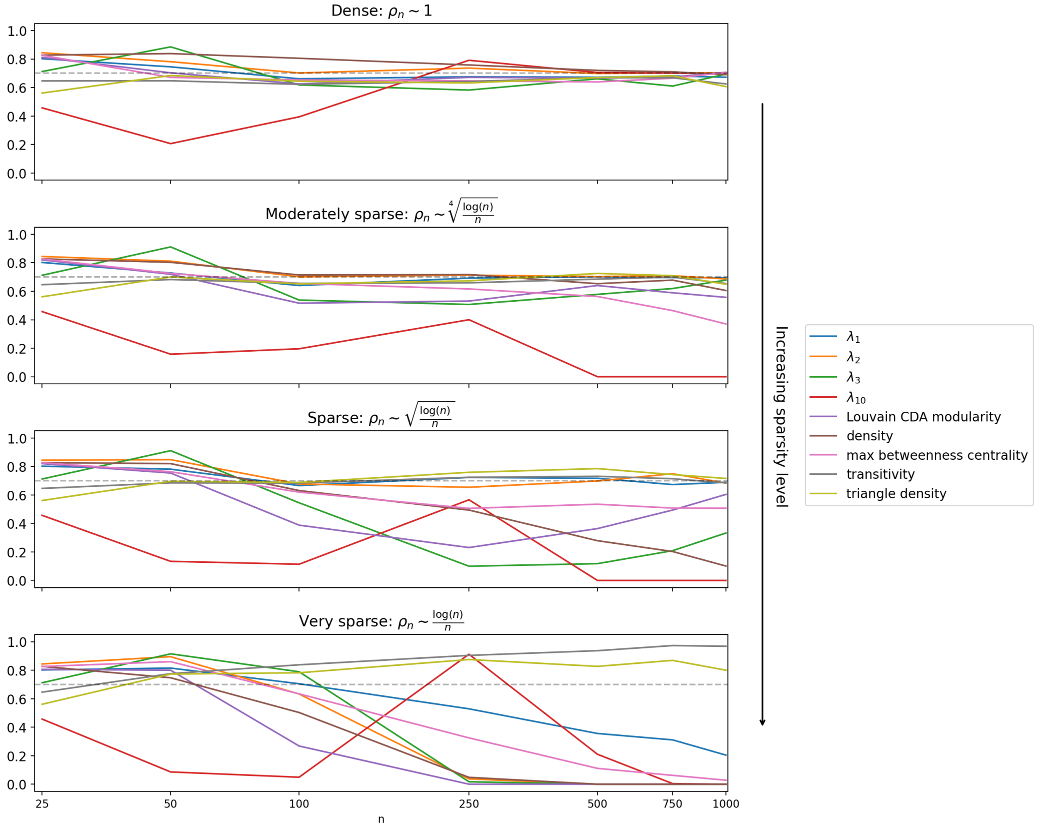
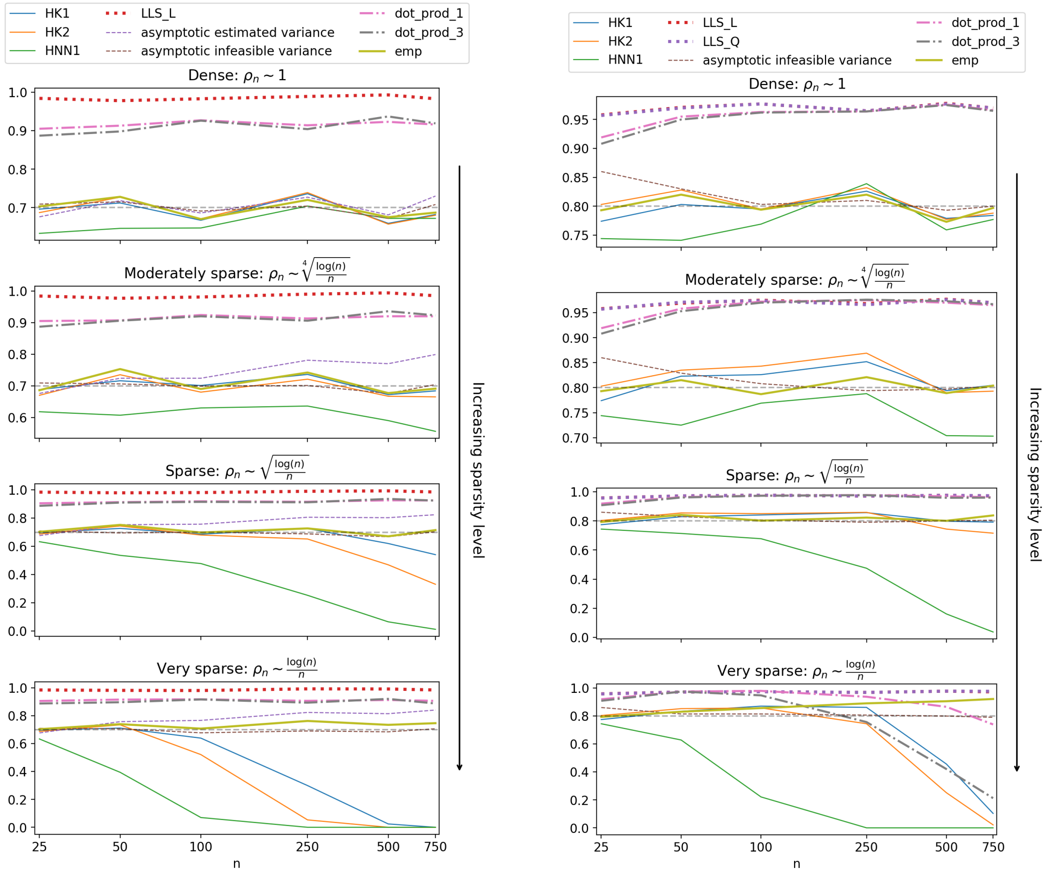


Figure 2.2: 70% confidence interval coverage for a range of statistics for the horseshoe generating function at different levels of sparsity. The sparsity levels are normalised to $\rho_n = 0.113$ at $n = 25$ and go down with n at the rates indicated above each subplot.

Next we compare the performance of different variations of our method and some of the existing competitor methods. Fig. 2.3a shows that variations of our method (HK1 and HK2) perform very similarly: we choose to use HK1 as the main method since HK2 is more computationally intensive and HK1 shows slight advantage for sparse graphs. Our method with Zhang, Levina, and Zhu (2017) linking function estimator (HNN1) performs slightly worse and its performance drops more significantly for sparser graphs. We believe this is mostly due to the choice of bandwidth (the estimator in Zhang, Levina, and Zhu (2017) relies on the theoretical optimal bandwidth instead of our numerically chosen $\hat{a}$). For U-statistics the empirical bootstrap of Green and Shalizi (2022) performs very well and remains good even at sparsity levels our methods cannot handle. This suggests that for sparse graphs we may want to consider lower bandwidth choice than $\hat{a}$, which would make our method more similar to the empirical bootstrap. The dot product bootstrap of Levin and Levina (2019) is presented for the correctly specified case of $k = 1$, as well as for $k = 3$. This method should have an advantage over the other ones since it is parametric and for $k = 1$ it is based on a correctly specified functional form of the linking function. However, it achieves coverage which is too high, over 90% instead of the required 70%. In simulations, we have seen that many estimates of linking probabilities ended up outside of the $[0, 1]$ region, which we believe is the reason why Levin and Levina (2019) estimators ended up biased. Their method should work better at larger sample sizes. The linear and quadratic methods from Lin, Lunde, and Sarkar (2020) are very similar to each other and suffer from the same issue of giving confidence intervals with higher coverage than desired.



(a) Density: 70% confidence interval coverage across different methods.

(b) Triangle density: 80% confidence interval coverage across different methods.

Figure 2.3: Confidence interval coverage for different methods using the product generating function. We compare: HK1 (our main method based on $\hat{h} \equiv \hat{h}^{(K1)}$ with $\hat{a}$), HK2 (our bootstrap method but using the linking function estimator $\hat{h}^{(K2)}$ with $a^{(optK2)}$ based on $\hat{h}^{(K2)}$), HNN1 (our bootstrap method but but using the linking function estimator $\hat{h}^{(NN1)}$ from Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size), emp (empirical bootstrap from Green and Shalizi (2022)), dot_prod_k (the bootstrap method from Levin and Levina (2019) based on assuming a k-dimensional ξ_i), asymptotic estimated variance (the asymptotic distribution from Bickel, Chen, and Levina (2011) with variance estimated according to the formula in Green and Shalizi (2022)), asymptotic infeasible variance (the asymptotic distribution from Bickel, Chen, and Levina (2011) with the true theoretical variance), LLS_L and LLS_Q (the linear and quadratic methods from Lin, Lunde, and Sarkar (2020)).

Most of the competitor methods can only be applied to U-statistics. An exception is the empirical bootstrap, which can be seen as a limiting case of our procedure with bandwidth close to 0. From Fig. 2.4 and Fig. 2.5 we can see that the empirical bootstrap would perform very poorly for statistics such as lower eigenvalues, max betweenness centrality and Louvain CDA modularity. There is also a version of the dot product bootstrap Levin and Levina (2019) which applies to more general functions, but, probably due to our bad coding or relatively small sample sizes, we were getting many estimated probabilities outside of $[0, 1]$ which led to very bad performance, likely not representative of the quality of their method, and is hence not presented here.

We have seen poor coverage for some of the more complicated statistics of interest, and in the final set of simulations we explore if this could be improved by choosing a different bandwidth than $\hat{a}$. We look at $c\hat{a}$ for a range of constants c. In Fig. 2.4 the confidence interval coverage for simple statistics such as density, triangle density or the largest eigenvalue is good at the default bandwidth and not too sensitive to the bandwidth choice: the performance remains good between $0.01\hat{a}$ to $4\hat{a}$. However, other statistics, such as eigenvalues below the largest one (e.g. λ_2, λ_3 and λ_{10} in Fig. 2.4) have poor coverage for the default choice of with $c = 1$, and low values of c. Luckily, in those cases we can fix the problem by choosing a wider bandwidth: the confidence interval coverage for $c \simeq 2$ remains very good. Hence for more complicated statistics, we need to be careful and check the performance not only at the default bandwidth choice but also at e.g. half or twice the default bandwidth.

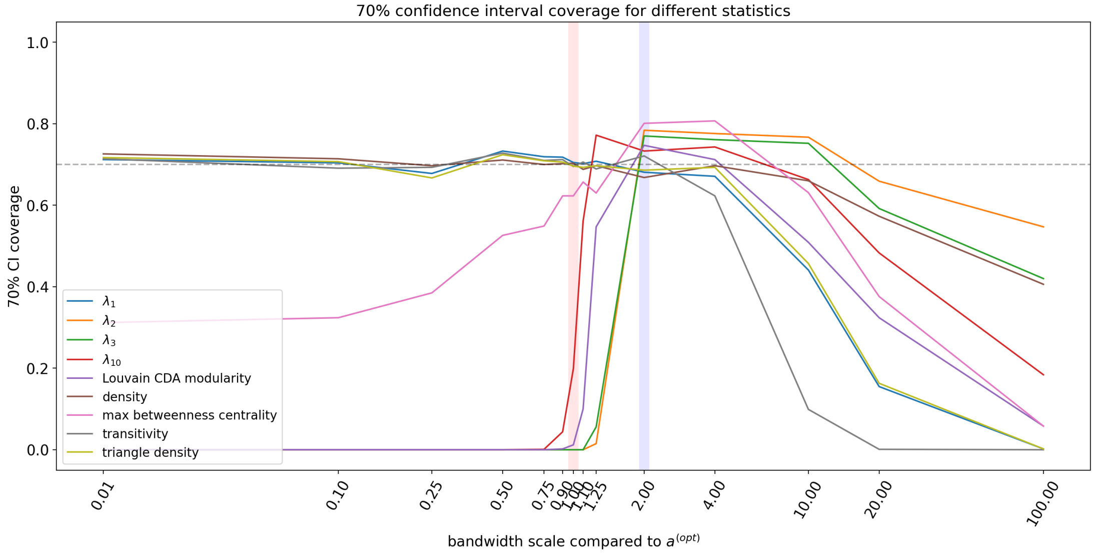


Figure 2.4: Confidence interval coverage for different bandwidths $c \times \hat{a}$: comparison of different statistics at $\alpha = 0.3$, $n = 300$ and $\rho_n = 0.1875$ based on Monte Carlo simulations using the product generating function.

These issues do not always arise: Fig. 2.5 shows an example when all statistics perform well for our default choice. For most statistics using a smaller bandwidth is not an issue: they do reasonably well for $c \leq 1$, even as small as $c = 0.01$, although they do not perform as well as for c close to 1. However, this is not universally true: some statistics, such as λ_{10} and maximal betweenness centrality, have very poor coverage outside of the region of $0.9 \leq c \leq 1.25$. The coverage for all statistics gets significantly worse when we use wider bandwidths ($c \geq 10$).

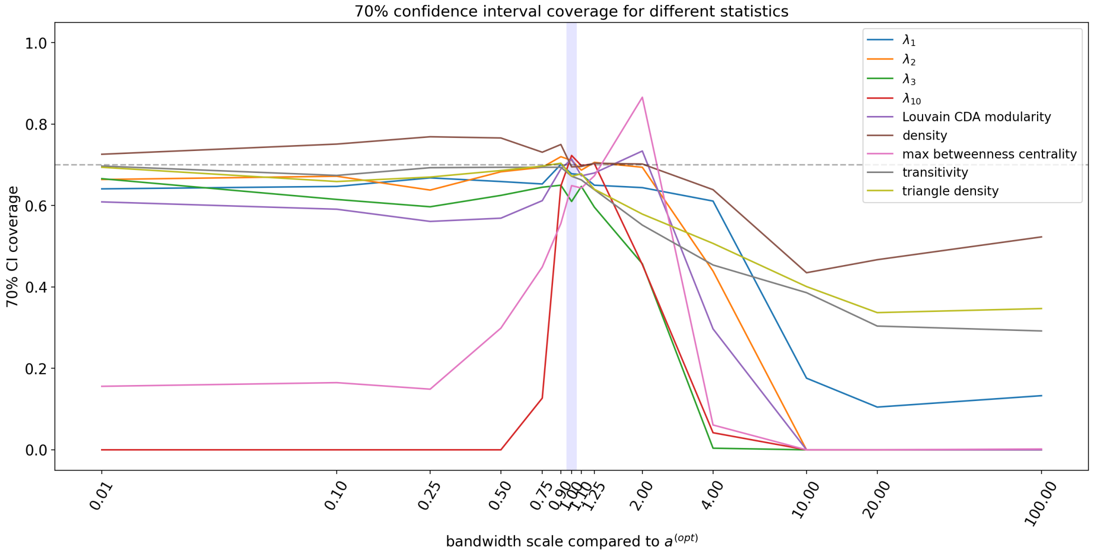
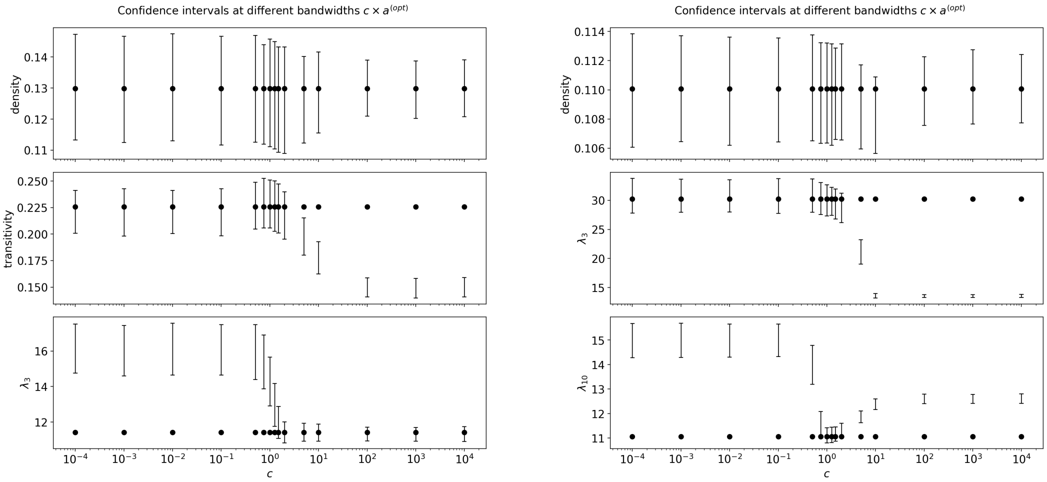


Figure 2.5: Confidence interval coverage for different bandwidths $c \times \hat{a}$: comparison of different statistics at $\alpha = 0.3$, $n = 500$ and $\rho_n = 0.1125$ based on Monte Carlo simulations using the horseshoe generating function.



(a) Confidence intervals for different statistics for the product generating function at $n = 300$ and $\rho_n = 0.125$.

(b) Confidence intervals for different statistics for the horseshoe generating function at $n = 500$ and $\rho_n = 0.1125$

Figure 2.6: Confidence intervals for different statistics and for bandwidths $c \times \hat{a}$ based on $B = 1000$ bootstrap graphs.

When the graph is sufficiently large and dense, c close to 1 gives good performance of most statistics we have checked. However, the performance does depend on the bandwidth choice and some statistics may be estimated poorly with the default bandwidth, especially when the graph is relatively sparse and the statistic is more complicated. In those cases, experimenting with different values of c could give more reliable results. This is one advantage of our method over the empirical bootstrap (which can be seen as a limiting case of our model with $c = 0$): while at the default setting both algorithms may be bad at estimating confidence intervals for lower eigenvalues, the performance of our method can be improved by selecting a larger bandwidth (oversmoothing) while the empirical bootstrap does not depend on any parameters that could be tweaked in a similar way.

The possibility of a poor performance at $\hat{a}$ raises a question: how can we choose the best bandwidth in an application, when we only have one observed network and no way to run a Monte Carlo simulation confirming the coverage? Luckily, there is an easy rule-of-thumb way to verify our choice. Fig. 2.6 shows an example of bootstrap confidence intervals formed from $B = 1000$ bootstrap replications for different statistics of a specific single true graph A estimated using different bandwidths. We can use it in the following way: if the statistic estimate from the original graph is in the middle of the confidence interval formed by bootstrapped graphs, the choice of the bandwidth is good. In Fig. 2.6a we see that density is always estimated relatively well, transitivity remains well-estimated for smaller than optimal bandwidths but is underestimated when the bandwidth is too large, and λ_3 is overestimated for smaller than optimal bandwidths but remains well estimated for larger than optimal bandwidth. In simulations, we have noticed a pattern that when the true value is above the estimated confidence interval lowering the bandwidth tends to improve the performance, while when the value from the original graph is below the confidence interval increasing the bandwidth often solves the problem. However, this is not always true: the lowest panel in Fig. 2.6b shows that λ_{10} is overestimated when bandwidth is either too low or too high compared to the optimal one. The middle panel also shows that the choice of a statistic does not determine the behaviour: for the horseshoe function λ_3 is better estimated for lower bandwidths and underestimated for higher ones, which is a different pattern than that of λ_3 from the product generating function in the bottom of Fig. 2.6a.

2.6 Application: the Diffusion of Microfinance

For our application, we use the data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013), a paper which analyses how information about microfinance spreads through social networks in 43 villages in India.

Prior to the introduction of a microfinance program they surveyed households in these villages and formed a network of connections based on 12 binary signals indicating if households knew each other (e.g. did they visit each other's homes, lend each other money, etc.). Each of these variables could be seen as a our A matrix, or we could combine them by taking a union, like in the original paper, to get an overall adjacency matrix. This is compatible with out framework: the closer two households are, the higher the probability that they will report each other as connected, hence we can view the reporting of a connection ($A_{ij} = 1$) as a signal from a Bernoulli distribution with probability of success proportional to the closeness of their friendship ($h_{0,n}(\xi_i, \xi_j)$).

Once the microfinance program entered the villages, they observed a set of first-informed villagers (injection points, chosen because they were village leaders who tend to be well-connected) and subsequent participation by households over a number of years.

One of the goals of the paper is to understand how the information about the program was spreading through the villages. This is modelled by a parametric diffusion model. Firstly, the probability p_{it} of household i with characteristics X_i participating when first informed is estimated from the logistic function:

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = X_i^{'}\beta.$$

The parameter $\hat{\beta}$ is estimated using the information about the leaders only. The aim is to estimate the probability of transferring information about a microfinance program by people who participate in it themselves (q^P) and by those who know about it but do not participate (q^N). This is done by simulating the information spreading over time discretised into T periods (trimesters). In each period the newly informed decide whether to participate ($m_{it} = 1$) with probability p_{it}, then each informed household spreads the information to its neighbours with probability $m_{it}q^P + (1 - m_{it})q^N$.

For each village v and for each set of discretised parameter values $\left(q^P, q^N\right)$ they simulate the spread of information and adoption decisions, and then calculate moments $m_{sim,v}\left(q^P, q^N\right)$ based on the final set of participating households (e.g. the fraction of households that have no participating neighbours but participate themselves, or the covariance of households participating in the program with the share of second neighbours that are participating). The average of these simulated moments across S simulations is compared to the observed empirical moments for the given village, $m_{emp,v}$. They then choose the parameter values which minimise the average of a function of deviation of simulated moments from empirical moments across all villages:

$$(\hat{q}^P, \hat{q}^N) =$$
$$= \arg\min_{q^P, q^N} \left(\frac{1}{43} \sum_{v=1}^{43} m_{sim,v}\left(q^P, q^N\right) - m_{emp,v} \right)' \hat{W} \left(\frac{1}{43} \sum_{v=1}^{43} m_{sim,v}\left(q^P, q^N\right) - m_{emp,v} \right).$$

where $\hat{W} = \frac{1}{43}\sum_{v=1}^{43}\left(m_{sim,v}\left(\tilde{q}^P, \tilde{q}^N\right) - m_{emp,v}\right)\left(m_{sim,v}\left(\tilde{q}^P, \tilde{q}^N\right) - m_{emp,v}\right)'$ for a first-stage estimates $\tilde{q}^P, \tilde{q}^N$ obtained by using I as the weighting matrix. To form confidence intervals they use bootstrap which resamples whole villages. The resulting estimates are shown as the "Original" ones in Fig. 2.7.

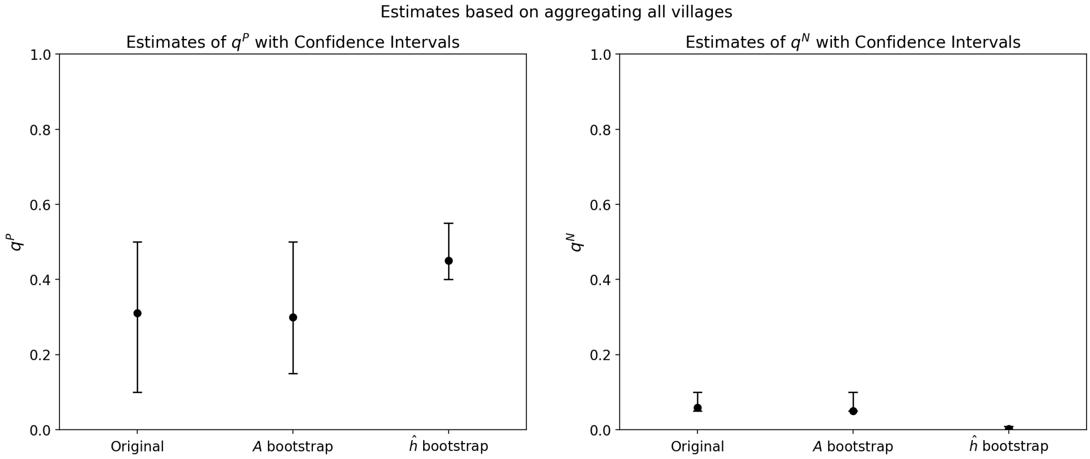


Figure 2.7: Estimates of q^P (left) and q^N (right) with 95% confidence intervals based on aggregating all villages: a comparison of the original result from Banerjee, Chandrasekhar, Duflo, and Jackson (2013) and our two methods.

The original paper considers a few variations of the model, including one which allows for endorsement effects. We only consider the information model without endorsement because it is less computationally demanding (the parameters are identified using a grid search and increasing the dimension of the parameter space by one leads to an exponential growth in the number of required simulations) and the original paper did not find evidence of a significant endorsement effect.

In our replication we use the same procedure for finding the parameters but we use our bootstrap to form confidence intervals. Instead of resampling whole villages we can estimate the matrix $\hat{h}_n$ for each of the villages[19] and use it to generate $B = 1000$ new sets of 43 villages with structures similar to the original ones. We can then repeat the whole estimation procedure for each new set of villages and obtain bootstrap estimates $(\hat{q}_b^{*P}, \hat{q}_b^{*N})$. The confidence intervals are formed by taking the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles for the distributions of $\hat{q}_b^{*P}$ and $\hat{q}_b^{*N}$. These

[19] The villages are assumed to be independent due to relatively large geographical distances between them. If they were not independent we could treat all households as belonging to one larger network.

estimates are presented in Fig. 2.7 as the "A bootstrap". They are very similar to the original estimates, with slightly narrower confidence intervals.[20]

The replication of the original result indicates that our method performs well, though it would not be advised in this situation because it is much more computationally demanding than the original bootstrap. However, with our setup we can do more. It is reasonable to assume that the spread of information is more likely between households which have a stronger connection (higher $h_{0,n}$), i.e. we can view the current model as an approximation to the true model in which the probability of spreading information depends not on the binary connection status A_{ij}, but on the actual strength of connection $h_{0,n}\left(\xi_i, \xi_j\right)$. Since $\hat{h}_n$ estimated as part of our procedure is a consistent estimate of $h_{0,n}$, we can repeat the simulations using a diffusion model based on $\hat{h}_n$ rather than on A (the imperfect signal about $h_{0,n}$). We assume that in any period $t \in \{1, \ldots, T\}$ the informed individual i spreads the information to another individual j with probability $\hat{h}_n\left(\xi_i, \xi_j\right)\left(m_{it}q^P + (1 - m_{it})q^N\right)$. The rest of the estimation procedure remains unchanged. The resulting estimates are reported as the "$\hat{h}$ bootstrap" estimates in Fig. 2.7. The confidence intervals are now much narrower than in the previous two cases and the conclusions differ as well: q^P is estimated to be higher (point estimate 0.45, 95% confidence interval $[0.4, 0.55]$) while q^N is essentially zero (point estimate 0.002, 95% confidence interval $[0, 0.009]$).

This contrasts with the findings of Banerjee, Chandrasekhar, Duflo, and Jackson (2013) who highlight the importance of non-participants in the diffusion process by showing that constraining q^N to be equal to zero leads to simulated participation dropping[21] from 20.0% to 13.97%. Our model shows that if we use a diffusion model based on $\hat{h}_n$ instead of A the estimated value of q^N is not distinguishable from zero and the simulated participation drops from 18.46% at the optimal values to 18.21% when we restrict q^N to 0, a drop of one seventy-fifth instead of one third. Using the more realistic assumption that the likelihood of spreading information depends on how well the households know each other removes the need for information spreading by non-participants.

Another extension made possible by our model is performing the analysis on individual village level. So far, we have assumed that the parameters are common across all villages, and with the original bootstrap resampling whole villages there was no way to form confidence intervals on at the village-level. With our bootstrap method, instead of minimising a (weighted) average of deviations of simulated moments from empirical moments across all villages, we can minimise them for each individual village. This allows us to:

1. **Check if all the villages come from the same network generating distribution.**

[20] Note that the confidence intervals for q^N cannot get much narrower because of the discretisation of the parameter space.

[21] The actual observed participation rate was 19.38%. It is not used as one of the moments matched in the parameter estimation.

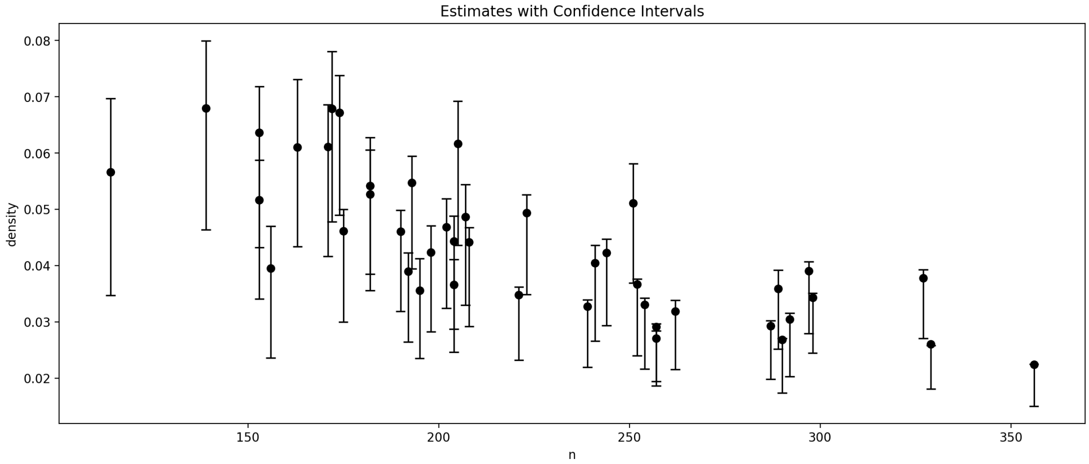


Figure 2.8: Densities of all 43 villages with bootstrap confidence intervals plotted against village size on the horizontal axis.

We bootstrap each village separately, find confidence intervals for some network statistics (e.g. density, or the largest eigenvalue) and see if these intervals overlap for all villages.

2. **Estimate the q^P and q^N parameters (and their confidence intervals) for each village separately, see if there are systematic differences between villages.** If there are differences, see if the hypotheses ($q^N > 0$ and $q^P > q^N$) hold in each village.

We firstly look at densities of all the 43 villages. Our bootstrap method allows us to not only obtain their point estimates but also add confidence intervals to see if the villages are systematically different from each other. Fig. 2.8 shows that the villages become more sparse as their size increases (consistent with the assumption that $\rho_n \to 0$ as $n \to \infty$). The confidence intervals in this graph are formed using the same bootstrapped villages that were used for estimating the model parameters.

Moving on to the model parameters, we have repeated the estimation using the original diffusion model based on the adjacency matrix A (Fig. 2.9) and the new diffusion model based on the linking probabilities $\hat{h}_n$ (Fig. 2.10). We can see that the two methods produce similar though not identical results. For some villages the estimation is very imprecise, leading to very wide confidence intervals. At least half of the villages have q^N precisely estimated to be zero, even in the model based on A: this may suggest that it is the imprecisely measured villages which drive the aggregate estimate to be positive.

In the last panels we can see that one of the predictions of the model, $q^P - q^N > 0$, cannot be concluded for most of the villages as we cannot reject the hypothesis that $q^P - q^N = 0$ (mostly due to imprecise measurements, in some part due to low numbers of participants in individual villages).

A practical extension of the current analysis would be to identify the village characteristics

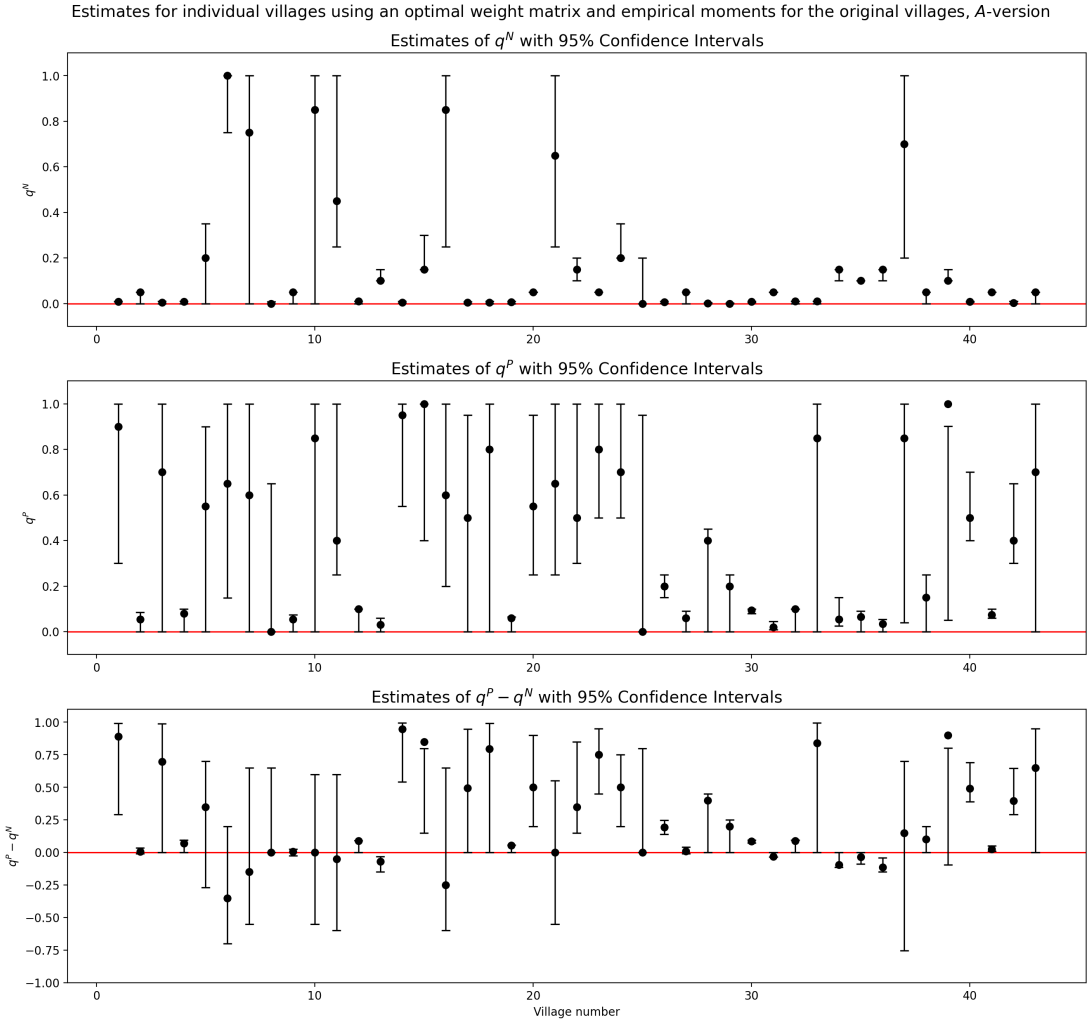


Figure 2.9: Estimates of q^N (top), q^P (middle) and $q^N - q^P$ (bottom) with 95% confidence intervals for individual villages using an optimal weight matrix and empirical moments for the original villages, A-version.

which help predict the estimated ranges of q^P and q^N. This would help policymakers choose villages in which the microfinance programs would have the highest chance of success or personalise the way in which the initial group of informed leaders is chosen depending on the information transmission characteristics in a given village.

2.7 Conclusion and Extensions

In this chapter we have proposed a network bootstrap procedure based on the nonparametric linking function estimator from Chapter 1 and we have provided conditions under which it consistently recovers the distribution of the original network and the distributions of a class of network functions related to U-statistics.

In the future projects we aim to provide a theoretical justification for consistency of our bootstrap method over a wider class of statistics, which is suggested by the promising results

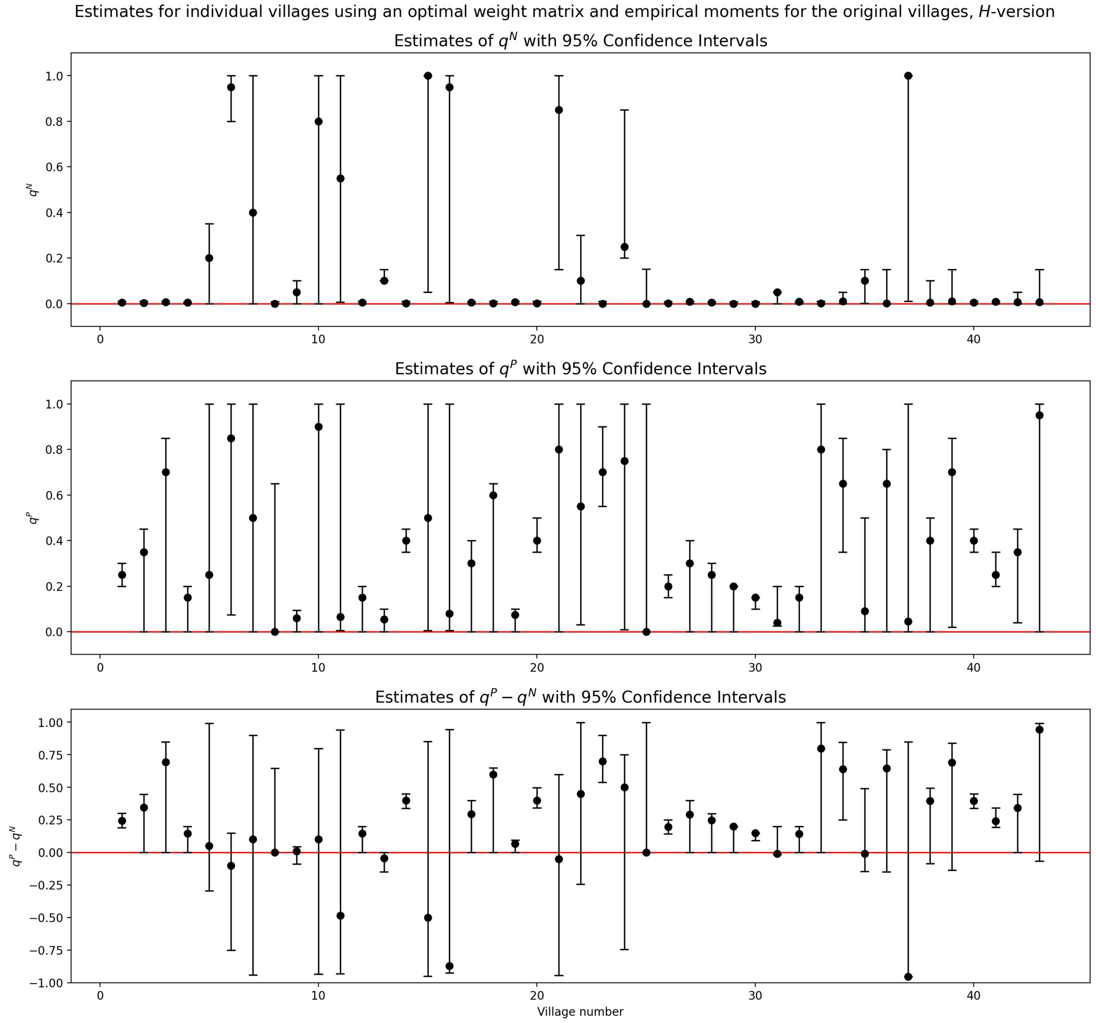


Figure 2.10: Estimates of q^N (top), q^P (middle) and $q^N - q^P$ (bottom) with 95% confidence intervals for individual villages using an optimal weight matrix and empirical moments for the original villages, H-version.

in our simulations. Most importantly, we would like to extend our results to regression models in which the outcome depends, possibly in a complicated way, on the entire adjacency matrix (e.g. spillover effects from neighbours). Unfortunately, it looks like in these cases the behaviour is not well approximated by that of an average of i.i.d. random variables, which makes deriving asymptotic results tricky. This is both an obstacle in proving bootstrap consistency and a reason why bootstrap methods are particularly needed when it comes to strongly dependent data structures such as networks.

One way in which we may be able to get around this issue is by looking for another notion of network distance than the Wasserstein metric proposed by Levin and Levina (2019). More specifically, one which would be sufficient for proving that the convergence is preserved after a transformation.

A less closely related future project inspired by this paper could be formulating a fast numerical procedure which allows for the selection of an objective-specific optimal bandwidth in two-step procedures. As in our setup, suppose we have a two-step estimation procedure where in the first step we estimate some parameter dependent on a tuning parameter (like $\hat{h}_n$ based on a_n), which we then plug into a (possibly random) second-step estimation (e.g. density in the resulting network). We ultimately care about the result of the second step (for us, the bootstrap confidence interval coverage of the statistic estimated in the second step) and we wish to optimise its performance by choosing an optimal tuning parameter in the first step. In this paper we have relied on optimising the first-step estimation ($\hat{a}$ was chosen to optimise the estimation of $\hat{h}_n$), but this was shown not to provide the best results in the second step across all second-step statistics. We have considered some algorithms which do rely on the second step performance (e.g. the approximate average MSE approach), but which are too slow to be useful in practice.

Bibliography for Chapter 2

Alatas, Vivi, Abhijit Banerjee, Arun G. Chandrasekhar, Rema Hanna, and Benjamin A. Olken. 2016. “Network Structure and the Aggregation of Information: Theory and Evidence from Indonesia.” *American Economic Review* 106 (7): 1663–1704.

Auerbach, Eric. 2022. “Identification and Estimation of a Partially Linear Regression Model Using Network Data.” *Econometrica* 90 (1): 347–365.

Banerjee, Abhijit, Emily Breza, Arun G Chandrasekhar, Esther Duflo, Matthew O Jackson, and Cynthia Kinnan. 2024. “Changes in social network structure in response to exposure to formal credit markets.” *Review of Economic Studies* 91 (3): 1331–1372.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2013. “The diffusion of microfinance.” *Science* 341 (6144): 1236498.

———. 2019. “Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials.” *The Review of Economic Studies* 86 (6): 2453–2490.

Bhattacharyya, Sharmodeep, and Peter J Bickel. 2015. “Subsampling bootstrap of count features of networks.” *The Annals of Statistics* 43 (6): 2384–2411.

Bickel, Peter J, Aiyou Chen, and Elizaveta Levina. 2011. “The method of moments and degree distributions for network models.”

Bickel, Peter J., and David A. Freedman. 1981. “Some Asymptotic Theory for the Bootstrap.” *The Annals of Statistics* 9 (6): 1196–1217.

Breza, Emily, and Arun G Chandrasekhar. 2019. “Social networks, reputation, and commitment: evidence from a savings monitors experiment.” *Econometrica* 87 (1): 175–216.

Chetty, Raj, Matthew O Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, et al. 2022. “Social capital I: measurement and associations with economic mobility.” *Nature* 608 (7921): 108–121.

Davezies, Laurent, Xavier D’Haultfœuille, and Yannick Guyonvarch. 2021. “Empirical process results for exchangeable arrays.” *The Annals of Statistics* 49 (2).

Giacomini, Raffaella, Dimitris N. Politis, and Halbert White. 2013. "A WARP-speed method for conducting monte carlo experiments involving bootstrap estimators." *Econometric Theory* 29 (3): 567–589.

Gine, Evarist, and Joel Zinn. 1990. "Bootstrapping General Empirical Measures." *The Annals of Probability* 18 (2): 851–869.

Green, Alden, and Cosma Rohilla Shalizi. 2022. "Bootstrapping exchangeable random graphs." *Electronic Journal of Statistics* 16 (1): 1058–1095.

Hahn, Jinyong. 1993. "Three essays in econometrics" [in English]. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-02-20. PhD diss.

Hansen, Bruce E. 2014. *Econometrics.* University of Wisconsin Department of Economics.

Levin, Keith, and Elizaveta Levina. 2019. "Bootstrapping networks with latent space structure." *arXiv preprint arXiv:1907.10821.*

Lin, Qiaohui, Robert Lunde, and Purnamrita Sarkar. 2020. "Trading off Accuracy for Speedup: Multiplier Bootstraps for Subgraph Counts." *arXiv preprint arXiv:2009.06170.*

Lunde, Robert, and Purnamrita Sarkar. 2022. "Subsampling sparse graphons under minimal assumptions." *Biometrika* 110 (1): 15–32.

Menzel, Konrad. 2021. "Bootstrap With Cluster-Dependence in Two or More Dimensions." *Econometrica* 89 (5): 2143–2188.

Politis, D.N., J.P. Romano, M. Wolf, P. Diggle, and S. Fienberg. 1999. *Subsampling.* Springer Series in Statistics. Springer New York.

Serfling, R.J. 2009. *Approximation Theorems of Mathematical Statistics.* Wiley Series in Probability and Statistics. Wiley.

Shao, Zhixuan, and Can M Le. 2024. "Parametric Bootstrap on Networks with Non Exchangeable Nodes." *arXiv preprint arXiv:2402.01866.*

Wang, Lawrence. 2016. "Network Comparisons using Sample Splitting." *PhD thesis, Carnegie Mellon University.*

Zeleneev, Andrei. 2020. "Identification and estimation of network models with nonparametric unobserved heterogeneity." *Department of Economics, Princeton University.*

Zhang, Yuan, Elizaveta Levina, and Ji Zhu. 2017. “Estimating network edge probabilities by neighbourhood smoothing.” *Biometrika* 104 (4): 771–783.

Zhang, Yuan, and Dong Xia. 2022. “Edgeworth expansions for network moments.” *The Annals of Statistics* 50 (2): 726–753.

Appendix

List of all notation

The notation in this file can get a bit heavy so we provide this list for reference.

- n – sample size, number of individuals in the network.
- A – an $n \times n$ adjacency matrix. Binary, symmetric, observed.
- A_{ij} – i, jth entry of the matrix A: 1 if i, j are connected (are neighbours), 0 if they are not.
- i, j, k, s, t – usually used to refer to one of the n individuals.
- ξ_i – vector of characteristics of individual i, enters the linking function.
- F_0 – distribution of ξ_i.
- $h_{0,n}$ – linking function, takes characteristics ξ_i, ξ_j as inputs and outputs the probability with which individuals i and j are linked. If the inputs are vectors $\xi(\iota) = (\xi_{\iota 1}, \xi_{\iota 2}, \ldots, \xi_{\iota m})$ of characteristics of multiple individuals it outputs the matrix of linking probabilities.
- ρ_n – density/sparsity parameter. Density in the sense that it is the expected edge density, sparsity in the sense that as $n \to \infty$ the density of edges decreases: $\rho_n \to 0$.
- w_0 – underlying linking probability before accounting for sparsity: $\rho_n w_0 = h_{0,n}$.
- $\varphi(\xi_i, \xi_t) = E\left(\frac{A_{is}A_{ts}}{\rho_n^2}|\xi_i, \xi_t\right)$ – a function measuring the probability of a common friend between i and j normalised by the sparsity level.
- $d_{ij} = \sqrt{E\left(E\left(w_0\left(\xi_t, \xi_s\right)\left(w_0\left(\xi_i, \xi_s\right) - w_0\left(\xi_j, \xi_s\right)\right)\right| \xi_i, \xi_j, \xi_t\right)^2\Big| \xi_i, \xi_j\Big)}$ – theoretical distance between i and j.
- $\hat{d}_{ij} = \frac{1}{\rho_n^2}\sqrt{\frac{1}{n}\sum_{t=1}^n\left(\frac{1}{n}\sum_{s=1}^n A_{ts}\left(A_{is} - A_{js}\right)\right)^2}$ – estimated distance between i and j.
- D_{ij} – shorthand notation for $\rho_n^4 \hat{d}_{ij}^2$ used in the description of the bootstrap procedure.

- $\hat{h}_n$ – estimated linking function:

$$\hat{h}_n(\xi_i, \xi_j) = \frac{\tilde{h}_n(\xi_i, \xi_j) + \tilde{h}_n(\xi_j, \xi_i)}{2} \quad \text{where} \quad \tilde{h}_n(\xi_i, \xi_j) = \frac{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right) A_{tj}}{\sum_{\substack{t=1 \\ t \neq j}}^{n} K\left(\frac{\rho_n^4 \hat{d}_{it}^2}{a_n}\right)}$$

- K – kernel function used in estimating linking probability.

- a_n – a bandwidth parameter, chosen by the researcher.

- $\hat{F}_n$ – the empirical distribution function of ξ_i; assigns equal probability to each of the original observations.

- * – a bootstrap equivalent, e.g. $\xi_i^* \sim \hat{F}_n$ is the bootstrap version of $\xi_i \sim F_0$.

- $\hat{}$ – an estimate.

- $\max_{i,j} \equiv \max_{i,j \in \{1,2,\ldots,n\}}$ – maximum over indices in a specific sample of size n.

- $\max_{\xi_i} \equiv \max_{\xi_i \in Supp(\xi_i)}$ – maximum over all $\xi_i \in Supp(\xi_i)$.

- $N(\xi_j, \delta) = \left\{\xi_k : \sup_{\xi_t} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta\right\}$ – the neighbourhood of ξ_j of size δ.

- $\omega(\delta) = \inf_{\xi_j \in Supp(\xi_j)} P\left(\xi_k \in N(\xi_j, \delta) | \, \xi_j\right)$ – the infimum over all possible ξ_i of the measures of their neighbourhoods of size δ.

- $b_n = \frac{a_n}{\rho_n^4}$ – a bandwidth parameter normalised by sparsity; the effective bandwidth size after accounting for the rate at which density goes to zero.

- $\hat{h}_n^-$ – leave-one-out version of $\hat{h}_n$, evaluated in the same way as $\hat{h}_n$ but without the observations $t = i, j$. Used for numerically choosing the optimal bandwidth.

- $\ell(A, a_n)$ – log-likelihood used for numerically choosing the optimal bandwidth. Defined in Eq. (1.12).

- $\hat{a}$ – numerically chosen optimal bandwidth. Defined in Eq. (1.13).

- B – number of bootstrap replications.

- $f_n(A_n(h_{0,n}(\xi), \eta), \rho_n, F_0)$ – a function whose distribution we are interested in.

- η – a vector of random variables which together with the linking function determine the realised links in A. We assume $\eta_{ij} \overset{ind}{\sim} \mathcal{U}[0,1]$ for $1 \leq i \leq j \leq n$ and η independent of ξ.

- $\tilde{f}_n\left(h_{0,n}\left(\xi\right), \rho_n, F\right) \equiv E(f_n(A_n(h_{0,n}(\xi), \eta), \rho_n, F_0)|\xi)$ – a function whose distribution we are interested in after averaging out the variation due to observing A instead of $h_{0,n}$.

- $E_{h_{0,n}}$, e.g. in $E_{h_{0,n}}(f_n(A_n(h_{0,n}(\xi),\eta),\rho_n,F_0)|\xi) = \int f_n(A_n(h_{0,n}(\xi),\eta),\rho_n,F_0)d\eta$ – expectation taken with respect to the independent Bernoulli trials with probabilities determined by $h_{0,n}$.

- $E_{h_{0,n},F_0}$ – expectation taken with respect to both the Bernoulli trials and the true distribution of ξ.

- ι – a vector of m nodes from $\{1,\dots,n\}$.

- $A(\iota)$ – the adjacency matrix of the subgraph with nodes ι (i.e. A from which we remove $n-m$ rows and columns not in ι).

- m – usually denotes the size of a subgraph or an order of U-statistic.

- g – a kernel function (in the U-statistic sense); a function of a subset of A.

- $\tilde{g}$ – a kernel function (in the U-statistic sense); a function of a subset of $h_{0,n}$. Equal to g after averaging out the variation due to observing A instead of $h_{0,n}$.

- $\tau(g)$ – a normalisation chosen to ensure $\frac{E_{h_{0,n},F_0}(g(A_n(\iota)))}{\rho_n^{\tau(g)}} = O_p(1)$.

- $\tilde{\tilde{g}}$ – the leading term in the normalised $\tilde{g}$; a function of a subset of w_0. $O_p(1)$.

- $\tilde{J}_n$ – the distribution of $\tilde{f}_n$.

- J_n – the distribution of f_n.

- $\hat{J}_{n,B}$ – an estimate of the distribution of f_n based on B bootstrap samples.

- J – limiting distribution of J_n as $n \xrightarrow{\infty}$: $J_n(t,h_{0,n},F_0) \Rightarrow J(t,w_0,F_0)$.

- $\Rightarrow$ – weak convergence.

- $\overset{a.s.}{\Rightarrow}$ – weak convergence almost surely, see Definition 2.4.3.

- $\overset{p}{\Rightarrow}$ – weak convergence in probability, see Definition 2.4.4.

- d_W – distance between measures which metrises weak convergence.

- $f(S) = \{f : S \to \mathbb{R} : |f(x)-f(y)| \leq d_S(x,y), \sup_{x\in S}|f(x)| \leq 1\}$ – the set of Lipschitz continuous and bounded real-valued functions on a metric space S equipped with distance d_S.

- $C_{w,F,\rho}$ – the set of non-random sequences of pairs of functions and distributions $\{(h_n,F_n)\}_{n=1}^{\infty}$ which satisfy a set of conditions on convergence of moments, see Definition 2.A.1.

- CI_n – a bootstrap confidence interval as defined in Eq. (2.22).

- $\mathcal{M}_m$ – the set of all possible multisets of cardinality m with elements from $\{1, 2, \ldots, m\}$

- $\hat{\rho}_n$ – estimator of density; the density of the observed adjacency matrix A.

- $\lambda_k(A)$ or λ_k – the kth largest eigenvalue of matrix A.

- $\hat{q}_\alpha$ – the estimate of αth quantile.

- q^P – in the application, the probability of transferring information about a microfinance program by program's participants

- q^N – in the application, the probability of transferring information about a microfinance program by those who do not participate the program themselves.

- C – generic positive constant, its value may change between different expressions in which it is used.

- C_ε – a positive constant which depends on $\varepsilon > 0$. Its value may change between different expressions in which it is used.

- T_n – a remainder term used in the proof of Theorem 1.

- M_w – an upper bound on the value of w_0: $\sup_{\xi_i, \xi_j} |w_0(\xi_i, \xi_j)| \leq M_w$.

- $r_n(i) = E\left(K\left(\frac{d_{it}^2}{b_n}\right)\middle|\,\xi_i\right)$ – the shorthand notation for the expected kernel weights based on the distance between i and other individuals used in the estimation of $\hat{h}_n\left(\xi_i, \xi_j\right)$.

- $\hat{r}_n(i) = \frac{1}{n-1}\sum_{\substack{t=1\\t\neq j}}^{n} K\left(\frac{\hat{d}_{it}^2}{b_n}\right)$ – the estimate of $r_n(i)$.

- $r_n = \inf_{\xi_i} r_n(i)$ – the smallest possible expected kernel weight. We need to ensure it is not too small or we would not be able to successfully estimate $h_{0,n}\left(\xi_i, \xi_j\right)$.

Appendix 2.A Proofs

Proof of Theorem 2.4.1. We start by constructing a particular coupling in $\Gamma\left(A^*, H\right)$. Let $\tilde{\gamma}$ be a particular joint distribution over $\hat{F}_n$ and F_0, the details of which we specify later in the proof. We use $\tilde{\gamma}$ to construct a coupling between A^* and H: we draw pairs $\{(\xi_i^*, \xi_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \tilde{\gamma}$. We also independently draw $\{\eta_{ij}\}_{i<j}^n \overset{i.i.d.}{\sim} \mathcal{U}[0,1]$ and set $\eta_{ij}^* = \eta_{ij}$. We denote $(A^*, H) \sim \tilde{\nu}$ and

note that this construction gives correct marginal distributions of A^* and H, hence:

$$\begin{aligned}
W_p^p\left(A^*, H\right) &\leq \int d_{GM}^p\left(A^*, H\right) d\tilde{\nu} \leq \int\left(\binom{n}{2}^{-1} \frac{\left\|A^*-H\right\|_{1,1}}{2}\right)^p d\tilde{\nu} \\
&\leq \int\binom{n}{2}^{-1} \sum_{i<j}\left|A_{ij}^*-H_{ij}\right|^p d\tilde{\nu}=\binom{n}{2}^{-1} \sum_{i<j} \int\left|A_{ij}^*-H_{ij}\right| d\tilde{\nu}=\int\left|A_{ij}^*-H_{ij}\right| d\tilde{\nu}
\end{aligned}$$

where the second inequality is due to the definition of d_{GM}, the third follows from the definition of $\frac{1}{2}\left\|A^*-H\right\|_{1,1}=\sum_{i<j}\left|A_{ij}^*-H_{ij}\right|$ and Jensen's inequality. The first equality is due to the fact that both adjacency matrices are binary ($1^p=1$, $0^p=0$) and the linearity of expectation. The final equality follows from the identity of distribution over all pairs (i, j). Expanding the final term:

$$\begin{aligned}
&\int\left|A_{ij}^*-H_{ij}\right| d\tilde{\nu}=\tilde{\nu}\left(\left\{A_{ij}^* \neq H_{ij}\right\}\right) \\
&\quad=\tilde{\nu}\left(\left\{\mathbb{1}\left(\hat{h}_n\left(\xi_i^*, \xi_j^*\right) \geq \eta_{ij}\right) \neq \mathbb{1}\left(h_{0,n}\left(\xi_i, \xi_j\right) \geq \eta_{ij}\right)\right\}\right) \\
&\quad=\int_0^1 \int \int\left|\mathbb{1}\left(\hat{h}_n\left(\xi_i^*, \xi_j^*\right) \geq \eta_{ij}\right)-\mathbb{1}\left(h_{0,n}\left(\xi_i, \xi_j\right) \geq \eta_{ij}\right)\right| d\tilde{\gamma}\left(\xi_i^*, \xi_i\right) d\tilde{\gamma}\left(\xi_j^*, \xi_j\right) d\eta_{ij} \\
&\quad=\int \int\left|\hat{h}_n\left(\xi_i^*, \xi_j^*\right)-h_{0,n}\left(\xi_i, \xi_j\right)\right| d\tilde{\gamma}\left(\xi_i^*, \xi_i\right) d\tilde{\gamma}\left(\xi_j^*, \xi_j\right) \\
&\quad\leq \int \int\left|\hat{h}_n\left(\xi_i^*, \xi_j^*\right)-h_{0,n}\left(\xi_i^*, \xi_j^*\right)\right| d\tilde{\gamma}\left(\xi_i^*, \xi_i\right) d\tilde{\gamma}\left(\xi_j^*, \xi_j\right) \\
&\qquad+\int \int\left|h_{0,n}\left(\xi_i^*, \xi_j^*\right)-h_{0,n}\left(\xi_i, \xi_j\right)\right| d\tilde{\gamma}\left(\xi_i^*, \xi_i\right) d\tilde{\gamma}\left(\xi_j^*, \xi_j\right)
\end{aligned}$$

The fourth equality follows from the fact that the two indicator functions differ in value only if η_{ij} falls into the interval between $h_{0,n}\left(\xi_i, \xi_j\right)$ and $\hat{h}_n\left(\xi_i^*, \xi_j^*\right)$, which happens with probability $\left|\hat{h}_n\left(\xi_i^*, \xi_j^*\right)-h_{0,n}\left(\xi_i, \xi_j\right)\right|$. In the last line we use triangle inequality.

We now look at the last two terms:

$$\begin{aligned}
&\int \int\left|\hat{h}_n\left(\xi_i^*, \xi_j^*\right)-h_{0,n}\left(\xi_i^*, \xi_j^*\right)\right| d\tilde{\gamma}\left(\xi_i^*, \xi_i\right) d\tilde{\gamma}\left(\xi_j^*, \xi_j\right) \\
&\quad=\int \int\left|\hat{h}_n\left(\xi_i^*, \xi_j^*\right)-h_{0,n}\left(\xi_i^*, \xi_j^*\right)\right| d\hat{F}_n\left(\xi_i^*\right) d\hat{F}_n\left(\xi_j^*\right) \\
&\quad=\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n\left|\hat{h}_n\left(\xi_i^A, \xi_j^A\right)-h_{0,n}\left(\xi_i^A, \xi_j^A\right)\right| \\
&\quad\leq \max_{i,j}\left|\hat{h}_n\left(\xi_i^A, \xi_j^A\right)-h_{0,n}\left(\xi_i^A, \xi_j^A\right)\right|=o_p(\rho_n)
\end{aligned}$$

by Theorem 1, where $\xi_i^A \sim F_0$ refers to the unobserved characteristics used in the formation of the matrix A which A^* is bootstrapped from.

For the other term we use Assumption 1.2, which says:

$$\inf_{\xi_j \in Supp(\xi_i)} P\left(\xi_k \in \left\{\sup_{\xi_t \in Supp(\xi_i)} |w_0(\xi_t, \xi_k) - w_0(\xi_t, \xi_j)| < \delta\right\}\right) \geq \left(\frac{\delta}{C}\right)^{\frac{1}{\alpha}}.$$

We have:

$$\begin{aligned}
&\int\int \left|h_{0,n}\left(\xi_i^*, \xi_j^*\right) - h_{0,n}\left(\xi_i, \xi_j\right)\right| d\tilde{\gamma}\left(\xi_i, \xi_i^*\right) d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) \\
&\quad \leq \rho_n \int\int \left|w_0\left(\xi_i^*, \xi_j^*\right) - w_0\left(\xi_i, \xi_j^*\right)\right| d\tilde{\gamma}\left(\xi_i, \xi_i^*\right) d\hat{F}_n\left(\xi_j^*\right) \\
&\qquad + \rho_n \int\int \left|w_0\left(\xi_i, \xi_j^*\right) - w_0\left(\xi_i, \xi_j\right)\right| dF_0\left(\xi_i\right) d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) \\
&\quad \leq \rho_n \int \sup_{\xi_j^* \in Supp(\xi_i)} \left|w_0\left(\xi_i^*, \xi_j^*\right) - w_0\left(\xi_i, \xi_j^*\right)\right| d\tilde{\gamma}\left(\xi_i, \xi_i^*\right) \\
&\qquad + \rho_n \int \sup_{\xi_i \in Supp(\xi_i)} \left|w_0\left(\xi_i, \xi_j^*\right) - w_0\left(\xi_i, \xi_j\right)\right| d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) \\
&\quad = 2\rho_n \int \sup_{\xi_i \in Supp(\xi_i)} \left|w_0\left(\xi_i, \xi_j^*\right) - w_0\left(\xi_i, \xi_j\right)\right| d\tilde{\gamma}\left(\xi_j, \xi_j^*\right)
\end{aligned}$$

The first inequality is due to the definition of $h_{0,n} = \rho_n w_0$ and triangle inequality. In the second inequality we take a supremum over the repeated index and note that the support of $\hat{F}_n$ is a subset of the support of F_0. As the terms no longer depend on ξ_j^* and ξ_i respectively, we integrate over their distributions. The resulting two terms are equal (note that w_0 is symmetric).

Fix $\varepsilon > 0$. For every $\xi_j^* \in Supp(\xi_i)$ there exists a neighbourhood $N(\xi_j^*, \varepsilon)$ of measure at least $\left(\frac{\varepsilon}{C}\right)^{\frac{1}{\alpha}}$ such that for all $\xi_j \in N(\xi_j^*, \varepsilon)$: $\sup_{\xi_i \in Supp(\xi_i)} \left|w_0\left(\xi_i, \xi_j^*\right) - w_0\left(\xi_i, \xi_j\right)\right| < \varepsilon$. Our task is to show that there exists a coupling $\tilde{\gamma}$ which aligns ξ_j^* with their corresponding neighbourhoods.

To that end, define

$$d_S(a, b) \equiv \sup_{\xi_i \in Supp(\xi_i)} \left|w_0\left(\xi_i, a\right) - w_0\left(\xi_i, b\right)\right|. \tag{2.24}$$

d_S is a pseudometric, i.e. it may fail positivity (the distance between two distinct points may be zero) but it satisfies all other properties of a distance (in particular the triangle inequality).

Take K points $\{a_1, \dots a_K\} \in Supp(\xi_i)$ which are at least ε apart: $\forall 1 \leq i < j \leq K$: $d_S(a_i, a_j) > \varepsilon$. Form a $\frac{\varepsilon}{2}$-neighbourhood around each a_k.

These neighbourhoods are non-overlapping: suppose there was a $b \in N\left(a_i, \frac{\varepsilon}{2}\right)$ and $b \in N\left(a_j, \frac{\varepsilon}{2}\right)$ for $i \neq j$. Then by triangle inequality: $d_S(a_i, a_j) \leq d_S(a_i, b) + d_S(a_j, b) \leq \varepsilon$. But we have assumed $d_S(a_i, a_j) > \varepsilon$, a contradiction.

By Assumption 1.2 we know that each of these neighbourhoods has a measure at least

$\left(\frac{\varepsilon}{2C}\right)^{\frac{1}{\alpha}}$. It follows that:

$$1 \geq P_{b\sim F_0}\left(b \in \bigcup_{i=1}^{K} N\left(a_i, \frac{\varepsilon}{2}\right)\right) = \bigcup_{i=1}^{K} P_{b\sim F_0}\left(b \in N\left(a_i, \frac{\varepsilon}{2}\right)\right) \geq K\left(\frac{\varepsilon}{2C}\right)^{\frac{1}{\alpha}}$$

or $K \leq \left(\frac{\varepsilon}{2C}\right)^{-\frac{1}{\alpha}} < \infty$, so the set of $\{a_1, \dots a_K\}$ has finite cardinality.

Take the largest K possible. Then for all $b \in Supp(\xi_i) \exists k \leq K$ such that $d_S(b, a_k) \leq \varepsilon$, or in other words $\bigcup_{i=1}^{K} N(a_i, \varepsilon)$ is a finite cover of $Supp(\xi_i)$. Hence we can assign each $b \in Supp(\xi_i)$ to one of the $k \in \{1, \dots, K\}$: start with $N\left(a_k, \frac{\varepsilon}{2}\right)$ for all k, then for each point not yet assigned to a region add it to the region with (not necessarily unique) k which minimises the d_S distance from that point to a_k. This way we form K disjoint regions, say $\{N_k\}_{k=1}^{K}$, each of size at least $\left(\frac{\varepsilon}{2C}\right)^{\frac{1}{\alpha}}$ and such that whenever $b_1, b_2 \in N_k \subseteq N(a_k, \varepsilon)$ we have $d_S(b_1, b_2) \leq d_S(b_1, a_k) + d_S(b_2, a_k) \leq 2\varepsilon$.

Now instead of ξ_i report $k(\xi_i)$ such that $\xi_i \in N_{k(\xi_i)}$. This means we are replacing F_0 with an empirical distribution function G_ε which takes only K values, each with probability $P_{b\sim F_0}(b \in N_k) \geq \left(\frac{\varepsilon}{2C}\right)^{\frac{1}{\alpha}}$; and we replace $\hat{F}_n$ with an empirical distribution function $\hat{G}_{\varepsilon,n}$ from G_ε. We choose $\tilde{\gamma}$ to be any coupling of $F_0, \hat{F}_n$ consistent with the following: for $y \sim \mathcal{U}[0,1]$ set $k(\xi_j) = G_\varepsilon^{-1}(y)$, $k\left(\xi_j^*\right) = \hat{G}_{\varepsilon,n}^{-1}(y)$.

Then:

$$\int \sup_{\xi_i \in Supp(\xi_i)} \left|w_0\left(\xi_i, \xi_j^*\right) - w_0\left(\xi_i, \xi_j\right)\right| d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) \leq 2\varepsilon + M_w \int \mathbb{1}\left(k(\xi_j) \neq k(\xi_j^*)\right) d\tilde{\gamma}\left(\xi_j, \xi_j^*\right)$$

For the first inequality we note that either ξ_j, ξ_j^* fall in the same N_k and hence $d_s\left(\xi_j^*, \xi_j\right) \leq 2\varepsilon$, or they come from different subsets of the domain, in which case their maximal possible distance is M_w. For the final term:

$$\begin{aligned}
\int \mathbb{1}\left(k(\xi_j) \neq k(\xi_j^*)\right) d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) &\leq \int \left|k(\xi_j) - k(\xi_j^*)\right| d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) \\
&= \int_0^1 \left|G_\varepsilon^{-1}(y) - \hat{G}_{\varepsilon,n}^{-1}(y)\right| dy \\
&= \int_1^K \left|G_\varepsilon(x) - \hat{G}_{\varepsilon,n}(x)\right| dx \\
&\leq K \sup_x \left|G_\varepsilon(x) - \hat{G}_{\varepsilon,n}(x)\right| \xrightarrow{a.s.} 0.
\end{aligned}$$

The first inequality is due to the fact that $k \in \mathbb{N}$ so if the terms are not equal their distance is at least 1. The next equality is by construction of $\tilde{\gamma}$, noting that $y \sim \mathcal{U}[0,1]$. We then do a change of variable (we switch from integrating the horizontal distance to the vertical distance between the plots of G_ε and $\hat{G}_{\varepsilon,n}$), noting that the plots can only differ on the domain

$x \in [1, K]$. We use an upper bound in terms of a supremum over x and conclude that the final expression goes to zero almost surely by Glivenko-Cantelli Theorem. Hence for all n large enough $\int \mathbb{1}\left(k(\xi_j) \neq k(\xi_j^*)\right) d\tilde{\gamma}\left(\xi_j, \xi_j^*\right) \leq \frac{\varepsilon}{M_w}$ with probability one.

Since ε was arbitrary, the overall expression is $o_p(\rho_n)$, as required. □

For the proofs of the next section, in the appendix we split the argument into more steps and provide intermediate results which lead to the conclusions in Theorem 2.4.2, Lemma 2.4.1 and Corollary 2.4.1. The advantage of the additional steps is that they characterise moment conditions sufficient for bootstrap consistency which could be verified for other classes of functions or alternative estimators of the network-generating function. They have been left out of the main text to avoid introducing more complicated notation and improve the readability.

We begin with a general result, not specific to U-statistics. Because the many levels of randomness can get confusing very quickly, we have decided to tackle them one at a time: we firstly characterise a class of *non-random* estimators and distributions for which we get weak convergence of our statistic to the correct limit. We denote these generic non-random statistics and distribution as e.g. h_n, F_n and we can think of them as specific realisations of their random equivalents, e.g. F_n can be the empirical distribution $\hat{F}_n|\xi$ we get for a specific draw of ξ. In practice, the classes of h_n, F_n will often be wider and also contain elements which cannot be achieved as a specific realisation of our random procedure. Once we have characterised the class which ensures weak convergence to the desired limit, we show that, once we allow for randomness in ξ, the statistics based on the random $\hat{h}_n, \hat{F}_n$ belong that class with high probability, hence they converge weakly to the same limit either almost surely or in probability.

Definition 2.A.1. *Set $C_{w,F,\rho}$. Let $\mathscr{h}$ denote a set of linking functions, let $\mathcal{F}$ denote a set of distributions, and let $(0,1]^{\mathbb{N}}$ denote a set of sequences of densities $\{\rho_n\}_{n=1}^{\infty}$, $0 < \rho_n \leq 1$. Let $(w, F, \rho) \in \mathscr{h} \times \mathcal{F} \times (0,1]^{\mathbb{N}}$ be a triple of a function w_0, a distribution F, and a sparsity sequence ρ. Let $\xi \sim F$ and $\xi^* \sim F_n$. For each $(w, F, \rho) \in \mathscr{h} \times \mathcal{F} \times (0,1]^{\mathbb{N}}$ let $C_{w,F,\rho}$ be the set of non-random sequences of pairs of functions and distributions $\{(h_n, F_n)\}_{n=1}^{\infty}$ characterised by a set of conditions on convergence of moments of the form $E_{F_n}\left(f\left(\frac{h_n}{\rho_n}(\xi^*), w(\xi^*)\right)\right) \to E_F\left(f\left(w(\xi), w(\xi)\right)\right)$ as $n \to \infty$ for some class of functions $f \in \mathscr{f}$. That is:*

$$
\begin{aligned}
C_{w,F,\rho} = \Big\{ & \{(h_n, F_n)\}_{n=1}^{\infty} : \forall n \in \mathbb{N}, \forall f \in \mathscr{f} : \\
& (h_n, F_n) \in \mathscr{h} \times \mathcal{F} \textit{ and } \lim_{n \to \infty} E_{F_n}\left(f\left(\frac{h_n}{\rho_n}(\xi^*), w(\xi^*)\right)\right) = E_F\left(f\left(w(\xi), w(\xi)\right)\right) \Big\}.
\end{aligned}
\tag{2.25}
$$

We state the general version of the result:[22]

Theorem 2.A.1. *Let $C_{w_0,F_0,\rho}$ be as defined in Definition 2.A.1 and suppose that:*

[22] The structure and the proof are strongly inspired by Theorem 1.2.1 of Politis et al. (1999).

(i) the set $C_{w_0,F_0,\rho}$ contains the sequence $\{(h_{0,n},F_0)\}_{n=1}^{\infty}$;

(ii) for any sequence $\{(h_n,F_n)\}_{n=1}^{\infty}$ in $C_{w_0,F_0,\rho}$, $\tilde{J}_n(t,h_n,F_n)$ converges weakly to a common distribution[23] $J(t,w_0,F_0)$;

(iii) for any sequence $\{(h_n,F_n)\}_{n=1}^{\infty}$ in $C_{w_0,F_0,\rho}$:

$$\lim_{n\to\infty} E_{h_n,F_n}\left[\left(f_n\left(A^*\left(h_n\left(\xi^*\right),\eta^*\right),\rho_n,F_n\right)-\tilde{f}_n(h_n\left(\xi^*\right),\rho_n,F_n)\right)^2\right]=0 \qquad (2.26)$$

where $A^*\left(h_n\left(\xi^*\right),\eta^*\right)$ denotes an adjacency matrix A^* based on a vector of observations of $\xi^* \overset{i.i.d.}{\sim} F_n$, with Bernoulli probabilities determined by $h_n\left(\xi_i^*\right)$.

If the random sequence $\{(\hat{h}_n,\hat{F}_n)\}_{n=1}^{\infty}$ belongs to $C_{w_0,F_0,\rho}$ with probability one, i.e. $\forall n\in\mathbb{N},\forall f\in f:(\hat{h}_n,\hat{F}_n)\in\hbar\times\mathcal{F}a.s.$ and $E_{\hat{F}_n}\left(f\left(\frac{\hat{h}_n}{\rho_n}(\xi^*),w_0(\xi^*)\right)\right)\xrightarrow{a.s.}E_{F_0}\left(f\left(w_0(\xi),w_0(\xi)\right)\right)$, then:

1. *$J_n(t,\hat{h}_n,\hat{F}_n)\overset{a.s.}{\Rightarrow}J(t,w_0,F_0)$.*

2. *If $J(t,w_0,F_0)$ is continuous in t at $t=1-\alpha$ and strictly increasing at $t=1-\alpha$:*

$$J_n^{-1}(1-\alpha,\hat{h}_n,\hat{F}_n)\xrightarrow{a.s.}J^{-1}(1-\alpha,w_0,F_0). \qquad (2.27)$$

3. *If $J(t,w_0,F_0)$ is continuous in t at $t=1-\alpha$ and is strictly increasing at $t=1-\alpha$ and if F_0 does not enter the function f_n directly but only through a parameter[24] θ: $f_n(A(h_{0,n}(\xi),\eta),\rho_n,\theta)$, then the $(1-\alpha)$ confidence interval for θ constructed as:*

$$CI_n\left(1-\alpha,A,\hat{h}_n,\hat{F}_n\right)=\left\{\theta:J_n^{-1}\left(\frac{\alpha}{2},\hat{h}_n,\hat{F}_n\right)\le f_n(A,\rho_n,\theta)\le J_n^{-1}\left(1-\frac{\alpha}{2},\hat{h}_n,\hat{F}_n\right)\right\} \qquad (2.28)$$

is asymptotically valid:

$$P_{h_{0,n},F_0}\left(\theta\in CI_n\left(1-\alpha,A,\hat{h}_n,\hat{F}_n\right)\right)\xrightarrow{a.s.}1-\alpha. \qquad (2.29)$$

4. *If $J\left(t,w_0,F_0\right)$ is continuous in t, then*

$$\sup_t\left|J_n\left(t,\hat{h}_n,\hat{F}_n\right)-\tilde{J}_n\left(t,h_{0,n},F_0\right)\right|\xrightarrow{a.s.}0.$$

If the random sequence $\{(\hat{h}_n,\hat{F}_n)\}_{n=1}^{\infty}$ satisfies the moment conditions for belonging to $C_{w_0,F_0,\rho}$ in probability: $E_{\hat{F}_n}\left(f\left(\frac{\hat{h}_n}{\rho_n}(\xi^*),w_0(\xi^*)\right)\right)\xrightarrow{p}E_{F_0}\left(f\left(w_0(\xi),w_0(\xi)\right)\right)$, then conclusions 1.-4. above hold with $\overset{p}{\Rightarrow}$ replacing $\overset{a.s.}{\Rightarrow}$ and $\xrightarrow{p}$ replacing $\xrightarrow{a.s.}$.

[23] This is weaker than F_n converges weakly to F_0.

[24] For example in equation (2.18) we have $\theta=E_{h_{0,n},F_0}(g(A(\iota)))$.

The above results follow straight from conveniently chosen assumptions, yet they are still useful because they provides a set of sufficient conditions for the convergence of the bootstrap distribution to the correct limit, the correctness of bootstrap confidence intervals, and the consistency of bootstrap.

Proof of Theorem 2.A.1. We start from proving 1.

For any $\{(h_n, F_n)\}_{n=1}^{\infty}$ in $C_{w_0,F_0,\rho}$:

$$\begin{aligned} f_n\left(A\left(h_n\left(\xi^*\right),\eta^*\right),\rho_n,F_n\right) &= \tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right) \\ &\quad + \Big(f_n\left(A\left(h_n\left(\xi^*\right),\eta^*\right),\rho_n,F_n\right) - \tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right)\Big) \end{aligned}$$

By assumption (ii), the distribution of $\tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right)$ converges weakly to the desired limit: $\tilde{J}_n(t,h_n,F_n) \xrightarrow{weakly} J(t,w_0,F_0)$. By assumption (iii), the second term converges to 0 in second mean, hence it is $o_p(1)$ and does not affect the distribution limit.[25] For any sequence $\{(h_n, F_n)\}_{n=1}^{\infty} \in C_{w_0,F_0,\rho}$ we have:

$$J_n(t,h_n,F_n) \xrightarrow{weakly} J(t,w_0,F_0) \quad \text{i.e.} \quad d\left(J_n(t,h_n,F_n), J(t,w_0,F_0)\right) \to 0.$$

The random sequence $\left\{(\hat{h}_n,\hat{F}_n)\right\}_{n=1}^{\infty}$ belongs to $C_{w_0,F_0,\rho}$ with probability one, hence:

$$\begin{aligned} &P\left(\lim_{n\to\infty} d\left(J_n(t,\hat{h}_n,\hat{F}_n), J(t,w_0,F_0)\right) = 0\right) \\ &\geq P\left(\{(\hat{h}_n,\hat{F}_n)\}_{n=1}^{\infty} \in C_{w_0,F_0,\rho} \text{ and } \lim_{n\to\infty} d\left(J_n(t,\hat{h}_n,\hat{F}_n), J(t,w_0,F_0)\right) = 0\right) \\ &= P\left(\{(\hat{h}_n,\hat{F}_n)\}_{n=1}^{\infty} \in C_{w_0,F_0,\rho}\right) \\ &= 1, \end{aligned}$$

that is: $d\left(J_n(t,\hat{h}_n,\hat{F}_n), J(t,w_0,F_0)\right) \xrightarrow{a.s.} 0$.

For the case of convergence in probability, we use the following result:

Theorem (Billingsley (1995) Theorem 20.5 (ii))**.** *A necessary and sufficient condition for $X_n \xrightarrow{p} X$ is that each subsequence $\{X_{n'}\}$ has a further subsequence $\{X_{n''}\}$ such that $X_{n''} \xrightarrow{a.s.} X$.*

Given that $E_{\hat{F}_n}\left(f\left(\frac{\hat{h}_n}{\rho_n}(\xi^*), w_0(\xi^*)\right)\right) \xrightarrow{p} E_{F_0}\left(f\left(w_0(\xi), w_0(\xi)\right)\right)$, for any subsequence in-

[25] By Theorem 25.4 in Billingsley (1995): $X_n \xrightarrow{d} X$ and $X_n - Y_n \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$. Also, if F_{X_n} and F_X denote the distribution functions of random variables X_n and X, respectively, then $X_n \xrightarrow{d} X$ means $F_{X_n} \xrightarrow{weakly} F_X$.

dexed by n' there is a further subsequence indexed by n'' which satisfies

$$E_{\hat{F}_{n''}}\left(f\left(\frac{\hat{h}_{n''}}{\rho_{n''}}(\xi^*), w_0(\xi^*)\right)\right) \xrightarrow{a.s.} E_{F_0}\left(f\left(w_0(\xi), w_0(\xi)\right)\right).$$

By what we have just shown, applied to $C_{w_0,F_0,\rho''}$, where ρ'' is the subsequence of ρ indexed by n'', $d\left(J_{n''}\left(t, \hat{h}_{n''}, \hat{F}_{n''}\right), J\left(t, w_0, F_0\right)\right) \xrightarrow{a.s.} 0$. Applying Theorem 20.5 (ii) from Billingsley (1995) in the other direction, this means that $d\left(J_n\left(t, \hat{h}_n, \hat{F}_n\right), J\left(t, w_0, F_0\right)\right) \xrightarrow{p} 0$. Hence 1. holds.

The remaining conclusions follow by arguments identical to those in the proof of Theorem 1.2.1 in Politis et al. (1999). For 2. we use the following Lemma:

Lemma (Lemma 1.2.1 of Politis et al. (1999))**.** *Let $\{G_n\}$ be a sequence of distribution functions on the real line converging weakly to a distribution function G (i.e. $G_n(x) \to G(x)$ for all continuity points of G). Assume G is continuous and strictly increasing at $y = G^{-1}(1-\alpha)$. Then,*

$$G_n^{-1}(1-\alpha) = \inf\{x : G_n(x) \geq 1-\alpha\} \to G^{-1}(1-\alpha). \tag{2.30}$$

Proof. See Politis et al. (1999) p.10. □

Together with the conclusion from 1. that $\tilde{J}_n(t, h_n, F_n) \xrightarrow{weakly} J(t, w_0, F_0)$ for all (h_n, F_n) in $C_{w_0,F_0,\rho}$, the lemma implies that $J_n^{-1}(1-\alpha, h_n, F_n) \to J^{-1}(1-\alpha, w_0, F_0)$ for all (h_n, F_n) in $C_{w_0,F_0,\rho}$. Arguments identical to those in the proof of 1. show that if $\{(\hat{h}_n, \hat{F}_n)\}_{n=1}^{\infty}$ belongs to $C_{w_0,F_0,\rho}$ with probability one, then $J_n^{-1}(1-\alpha, \hat{h}_n, \hat{F}_n) \xrightarrow{a.s.} J^{-1}(1-\alpha, w_0, F_0)$ and if $\{(\hat{h}_n, \hat{F}_n)\}_{n=1}^{\infty}$ satisfies the moment conditions for belonging to $C_{w_0,F_0,\rho}$ in probability: $E_{\hat{F}_n}\left(f\left(\frac{\hat{h}_n}{\rho_n}(\xi^*), w_0(\xi^*)\right)\right) \xrightarrow{p} E_{F_0}\left(f\left(w_0(\xi), w_0(\xi)\right)\right)$, then $J_n^{-1}(1-\alpha, \hat{h}_n, \hat{F}_n) \xrightarrow{p} J^{-1}(1-\alpha, w_0, F_0)$.

In order to show 3., we firstly prove the following Lemma:

Lemma 2.A.1. *Let $\{G_n\}$ be a sequence of distribution functions on the real line converging weakly to a distribution function G (i.e. $G_n(x) \to G(x)$ for all continuity points of G). Let x_n be a real-valued sequence converging to x (i.e. $x_n \to x$). Assume that G is continuous and strictly increasing at x. Then,*

$$G_n(x_n) \to G(x). \tag{2.31}$$

Proof. Take any $\delta > 0$. Since G is continuous at x, there exists $\varepsilon > 0$ such that $x - \varepsilon$ and $x + \varepsilon$

are continuity points of G and

$$\begin{aligned} G(x-\varepsilon) - G(x) &\geq -\frac{\delta}{2} \\ G(x+\varepsilon) - G(x) &\leq \frac{\delta}{2}. \end{aligned}$$

Since $x_n \to x$, $G_n(x-\varepsilon) \to G(x-\varepsilon)$, $G_n(x) \to G(x)$ and $G_n(x+\varepsilon) \to G(x+\varepsilon)$, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$:

$$\begin{aligned} |x_n - x| &\leq \varepsilon \\ |G_n(x-\varepsilon) - G(x-\varepsilon)| &\leq \frac{\delta}{2} \\ |G_n(x) - G(x)| &\leq \frac{\delta}{2} \\ |G_n(x+\varepsilon) - G(x+\varepsilon)| &\leq \frac{\delta}{2}. \end{aligned}$$

Since G_n are weakly increasing for all n:

$$G_n(x-\varepsilon) \leq G_n(x) \leq G_n(x+\varepsilon).$$

Hence for all $n \geq N$:

$$\begin{aligned} -\delta \leq G(x-\varepsilon) - G(x) - \frac{\delta}{2} &\leq G_n(x-\varepsilon) - G(x) \leq \\ &\leq G_n(x) - G(x) \leq \\ &\leq G_n(x+\varepsilon) - G(x) \leq G(x+\varepsilon) - G(x) + \frac{\delta}{2} \leq \delta. \end{aligned}$$

i.e. $|G_n(x_n) - G(x)| \leq \delta$. □

For 3, we start with any (h_n, F_n) in $C_{w_0,F_0,\rho}$. We have:

$$\begin{aligned} &P_{h_{0,n},F_0}\left(\theta \in CI_n\left(1-\alpha, A, h_n, F_n\right)\right) \\ &\quad = P_{h_{0,n},F_0}\left(J_n^{-1}\left(\frac{\alpha}{2}, h_n, F_n\right) \leq f_n(A,\rho_n,\theta) \leq J_n^{-1}\left(1-\frac{\alpha}{2}, h_n, F_n\right)\right) \\ &\quad = P_{h_{0,n},F_0}\left(f_n(A,\rho_n,\theta) \leq J_n^{-1}\left(1-\frac{\alpha}{2}, h_n, F_n\right)\right) \\ &\qquad\quad - P_{h_{0,n},F_0}\left(f_n(A,\rho_n,\theta) < J_n^{-1}\left(\frac{\alpha}{2}, h_n, F_n\right)\right) \\ &\quad = J_n\left(J_n^{-1}\left(1-\frac{\alpha}{2}, h_n, F_n\right), h_{0,n}, F_0\right) - J_n\left(J_n^{-1}\left(\frac{\alpha}{2}, h_n, F_n\right), h_{0,n}, F_0\right) \\ &\quad \to J\left(J^{-1}\left(1-\frac{\alpha}{2}, w_0, F_0\right), w_0, F_0\right) - J\left(J^{-1}\left(\frac{\alpha}{2}, w_0, F_0\right), w_0, F_0\right) = 1-\alpha. \end{aligned}$$

The convergence follows from Lemma 2.A.1 used with 2. (for the convergence of the argument)

and $J_n(t, h_{0,n}, F_0) \xrightarrow{weakly} J(t, w_0, F_0)$ (for the convergence in distribution). Arguments identical to those in the proof of 1. show that if $\left\{\left(\hat{h}_n, \hat{F}_n\right)\right\}_{n=1}^{\infty}$ belongs to $C_{w_0,F_0,\rho}$ with probability one, then $P_{h_{0,n},F_0}\left(\theta \in CI_n\left(1-\alpha, A, \hat{h}_n, \hat{F}_n\right)\right) \xrightarrow{a.s.} 1-\alpha$. If instead $\left\{\left(\hat{h}_n, \hat{F}_n\right)\right\}_{n=1}^{\infty}$ satisfies the moment conditions for belonging to $C_{w_0,F_0,\rho}$ only in probability, that is:
$E_{\hat{F}_n}\left(f\left(\frac{\hat{h}_n}{\rho_n}(\xi^*), w_0(\xi^*)\right)\right) \xrightarrow{p} E_{F_0}\left(f\left(w_0(\xi), w_0(\xi)\right)\right)$, then we have
$P_{h_{0,n},F_0}\left(\theta \in CI_n\left(1-\alpha, A, \hat{h}_n, \hat{F}_n\right)\right) \xrightarrow{p} 1-\alpha$.

Finally, 4. follows from 1. and Polya's Theorem:

Theorem (Polya's Theorem, Satz I of Pólya (1920))**.** *Let X_n, X be random variables with distributions $F_n(x)$ and $F(x)$ respectively. If F is continuous*

$$X_n \xrightarrow{d} X \iff \sup_x |F_n(x) - F(x)| \to 0.$$

□

We can now provide more primitive conditions for the special class of f_n for which $\tilde{f}_n$ is a U-statistic.

Theorem 2.A.2 (Consistency of bootstrap for U-statistics)**.** *Let ι be a set of m nodes and denote the adjacency matrix on the subgraph with nodes in ι and linking probabilities $h_n(.,.)$ by $A(h_n(\xi(\iota)), \eta(\iota))$. Let $g : \{0,1\}^{\binom{m}{2}} \to \mathbb{R}$ be a symmetric function from a subgraph on $m < \infty$ nodes to the real line and let*

$$\begin{aligned} & f_n(A(h_n(\xi^*), \eta^*), \rho_n, F_n) \\ & \quad = \frac{\sqrt{n}}{\binom{n}{m}\rho_n^{\tau(g)}} \sum_{1 \le \iota_1 < \iota_2 < \cdots < \iota_m \le n} \left(g(A(h_n(\xi^*(\iota)), \eta^*(\iota))) - E_{h_n,F_n}(g(A(h_n(\xi^*(\iota)), \eta^*(\iota))))\right). \end{aligned}$$

and $\tilde{g}(h_{0,n}(\xi(\iota))) \equiv E(g(A(h_{0,n}(\xi(\iota)), \eta(\iota)))|\xi(\iota))$. There exists a normalisation[26] *$\tau(g)$ and a function $\tilde{\tilde{g}} : Supp(\xi)^m \to \mathbb{R}$ such that:*

- $\frac{\tilde{g}(h_{0,n}(\xi(\iota)))}{\rho_n^{\tau(g)}} = \tilde{\tilde{g}}(w_0(\xi(\iota))) + O(\rho_n)$
- $E_{F_0}(|\tilde{\tilde{g}}(w_0(\xi(j)))|) > 0$ *for some* $j \in \mathcal{M}_m$
- $E_{F_0}\left(\tilde{\tilde{g}}^2\left(w_0\left(\xi(j)\right)\right)\right) < \infty \quad \forall j \in \mathcal{M}_m$
- $Var_{F_0}(E_{F_0}(\tilde{\tilde{g}}(w_0(\xi(\iota)))|\xi_{\iota_1})) \equiv \sigma_1^2 < \infty$

[26] For $m = 2$, if $g(0) \neq 0$ we set $\rho_n^{-\tau(g)} = 1$, $\tilde{\tilde{g}}(w_0(\xi_i, \xi_j)) = g(0)$ and if $g(0) = 0$ but $g(1) \neq 0$ we set $\rho_n^{-\tau(g)} = \frac{1}{\rho_n}$ and $\tilde{\tilde{g}}(w_0(\xi_i, \xi_j)) = g(1)w_0(\xi_i, \xi_j)$. More generally, for $m \geq 2$, $\rho_n^{-\tau(g)} = \frac{1}{\rho_n^k}$ where k is the smallest number of ones such that $g(\cdot)$ evaluated at a vector of k ones and $\binom{m}{2} - k$ zeros is non-zero.

Suppose that:

$$\sigma_1^2 > 0$$
$$\frac{n}{\binom{n}{m}\rho_n^{\tau(g)}} \to 0.$$

and let $C_{w_0,F_0,\rho}$ be a set of sequences $\{(h_n, F_n)\}_{n=1}^{\infty}$ which satisfy:

1. $E_{F_n}\left(\left(\frac{1}{\rho_n}\left(h_n(\xi_i^*, \xi_j^*) - h_{0,n}(\xi_i^*, \xi_j^*)\right)\right)^2\right) \to 0.$

2. $E_{F_n}\left(f\left(\xi^*(\iota)\right)\right) \to E_{F_0}\left(f\left(\xi(\iota)\right)\right)$ *for all* $f : Supp(\xi)^k \to \mathbb{R}$ *such that* $E_{F_0}\left(\left|f\left(\xi(\iota)\right)\right|\right) < \infty$ *for all* $\iota \in \mathcal{M}_k$*, for any* $k \leq 2m - 1$.

Then $\{(\hat{h}_n, \hat{F})\}_{n=1}^{\infty}$ satisfies 1.-2. in probability and we get all conclusions of Theorem 2.A.1 in probability with $J(t, w_0, F_0) = N(0, m^2\sigma_1^2)$.

Remark. *The advantage of stating our condition as in 2. instead of directly showing that it holds when $F_n = \hat{F}_n$ because of SLLN for U-statistics is that it characterises a wider class of distributions we could resample from. For example, when we adjust the resampling distribution for the purpose of GMM by adding weights to different observations in a way that ensures the moment conditions hold in the bootstrap world.*

Proof of Theorem 2.A.2. The theorem was stated for a general $m < \infty$ but for simplicity of notation we present the proof for the case of $m = 2$. The structure of the argument remains identical if we use $m > 2$.

To show the existence of $\tilde{\tilde{g}}$ and $\tau(g)$ we start by analysing the form of $\tilde{g}$. Since g is a function from $\{0,1\}^{\binom{m}{2}}$ it takes at most $2^{\binom{m}{2}}$ distinct values. Each of those values is taken with probability that the input submatrix $A(\iota)$ matches a given pattern of 0s and 1s. Let $\Gamma(A(\iota))$ denote the set (of cardinality $2^{\binom{m}{2}}$) of all possible values $A(\iota)$ can take. Then:

$$\tilde{g}(h_{0,n}(\xi(\iota))) = \sum_{\gamma \in \Gamma(A(\iota))} g(\gamma) P(A(\iota) = \gamma | \xi(\iota)).$$

Conditional on $\xi(\iota)$, the elements of $A(\iota)$ are independent and $P(A_{ij} = 1|\xi) = h_{0,n}(\xi_i, \xi_j) = \rho_n w_0(\xi_i, \xi_j) \sim \rho_n$ while $P(A_{ij} = 0|\xi) = 1 - h_{0,n}(\xi_i, \xi_j) = 1 - \rho_n w_0(\xi_i, \xi_j) \sim 1$. The probability of the event that the upper triangle of $A(\iota)$ consists of k ones and $\binom{m}{2} - k$ zeros is proportional to ρ_n^k. The smallest k for which $g(\cdot)$ evaluated at an input γ with k ones and $\binom{m}{2} - k$ zeros in the upper triangle is non-zero is equal to the normalisation $\tau(g)$. By construction, all γs with fewer ones have a coefficient $g(\gamma) = 0$. All γs with more ones happen with probability proportional to ρ_n^l for $l > \tau(g)$, i.e. after a normalisation by $\rho_n^{-\tau(g)}$ are $O\left(\rho_n\right)$ and go to zero.

The terms in the sum proportional to $\rho_n^{\tau(g)}$ are of the form:

$$g(\gamma)\underbrace{h_{0,n}(\xi_{\iota_1},\xi_{\iota_2})\dots h_{0,n}(\xi_{\iota_3},\xi_{\iota_4})}_{\tau(g)\text{ terms}}(1-h_{0,n}(\xi_{\iota_5},\xi_{\iota_6}))\dots(1-h_{0,n}(\xi_{\iota_7},\xi_{\iota_8})).$$

After a normalisation by $\rho_n^{-\tau(g)}$ we get:

$$g(\gamma)\underbrace{w_0(\xi_{\iota_1},\xi_{\iota_2})\dots w_0(\xi_{\iota_3},\xi_{\iota_4})}_{\tau(g)\text{ terms}}(1-h_{0,n}(\xi_{\iota_5},\xi_{\iota_6}))\dots(1-h_{0,n}(\xi_{\iota_7},\xi_{\iota_8}))$$

we keep the $g(\gamma)w_0(\xi_{\iota_1},\xi_{\iota_2})\dots w_0(\xi_{\iota_3},\xi_{\iota_4})$ part in $\tilde{\tilde{g}}$ and note that the remainder of the previous term is $O(\rho_n)$.

To sum up, $\tilde{\tilde{g}}$ takes the form of a finite sum of non-zero constant (value of g at a specific realisation γ) times a product of $\tau(g)$ terms of the form $w_0(\xi_{\iota_i},\xi_{\iota_j})$.

The remaining terms in $\frac{\tilde{g}(h_{0,n}(\xi(\iota)))}{\rho_n^{\tau(g)}} - \tilde{\tilde{g}}(w_0(\xi(\iota)))$ vanish at the rate $O(\rho_n)$.

Since $w_0(\xi_i,\xi_j)$ is not identically equal to zero and there are non-zero coefficients $g(\gamma)$ multiplying products of $w_0(\xi_{\iota_i},\xi_{\iota_j})$ in $\tilde{\tilde{g}}$, there exists $j\in\mathcal{M}_m$ for which $E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi(j)))|\right)>0$.

Since $w_0(\xi_i,\xi_j)<M_w$ for all ξ_i,ξ_j we have

$$E_{F_0}\left(\tilde{\tilde{g}}^2\left(w_0\left(\xi(j)\right)\right)\right)<\binom{\binom{m}{2}}{\tau(g)}^2\left(\max_{\gamma\in\Gamma(A(\iota))}g^2(\gamma)\right)M_w^2<\infty$$

for all $j\in\mathcal{M}_m$ (where the first constant says that there are $\binom{m}{2}$ ones and zeros that determine the value of $A(\iota)$, there are $\binom{\binom{m}{2}}{\tau(g)}$ ways to place $\tau(g)$ ones in them, and after squaring a sum of $\binom{\binom{m}{2}}{\tau(g)}$ terms we get $\binom{\binom{m}{2}}{\tau(g)}^2$ terms, each bounded above by the remaining part of the expression). Finally, since

$$Var(E(Y|X))=Var(Y)-E(Var(Y|X))\le Var(Y)=E(Y^2)-E(Y)^2\le E(Y^2)$$

we also get that

$$\sigma_1^2\equiv Var_{F_0}(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi(\iota)))|\xi_{\iota_1}\right))<E_{F_0}\left(\tilde{\tilde{g}}^2\left(w_0\left(\xi(\iota)\right)\right)\right)<\infty.$$

Having established the existence of $\tilde{\tilde{g}}$ and $\tau(g)$, we now check that the elements of $C_{w_0,F_0,\rho}$ satisfy condition (i)-(iii) of Theorem 2.A.1.

The sequence $\{(h_n,F_n)\}_{n=1}^{\infty}=\{(h_{0,n},F_0)\}_{n=1}^{\infty}$ satisfies the conditions and belongs to $C_{w_0,F_0,\rho}$ (sequences in 1. and 2. are constant and equal to the desired limit), hence (i) is satisfied.

To check condition (iii) we look at:

$$
\begin{aligned}
&E_{h_n,F_n}\left(\left(f_n(A^*\left(h_n\left(\xi^*\right),\eta^*\right),\rho_n,F_n) - \tilde{f}_n(h_n\left(\xi^*\right),\rho_n,F_n)\right)^2\right) \\
&= \frac{n}{\binom{n}{2}^2\rho_n^{2\tau(g)}}\sum_{i^*<j^*}\sum_{k^*<l^*} E_{h_n,F_n}\left(\left(g(A_{i^*j^*}) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)\left(g(A_{k^*l^*}) - E_{h_n}\left(g(A_{k^*l^*})|\xi^*\right)\right)\right)
\end{aligned}
$$

To simplify the above expression notice that most terms in the summation are zero. In particular, consider different cases of overlap between the indices:

- if there is no overlap ($i^* \neq k^*$, $i^* \neq l^*$, $j^* \neq k^*$, $j^* \neq l^*$), by the independence assumption the term inside the sum is:

$$
E_{h_n,F_n}\left(\left(g(A_{i^*j^*}) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)\right)^2 = 0^2 = 0.
$$

- If there is partial overlap (e.g. $i^* = k^*$, $j^* \neq l^*$, or any symmetric situation):

$$
\begin{aligned}
&E_{h_n,F_n}\left(\left(g(A_{i^*j^*}) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)\left(g(A_{i^*l^*}) - E_{h_n}\left(g(A_{i^*l^*})|\xi^*\right)\right)\right) \\
&\overset{LIE}{=} E_{h_n,F_n}\left(E_{h_n,F_n}\left(\left(g(A_{i^*j^*}) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)\left(g(A_{i^*l^*}) - E_{h_n}\left(g(A_{i^*l^*})|\xi^*\right)\right)|\xi_i^*\right)\right) \\
&\overset{indep}{=} E_{h_n,F_n}\left(\left(E_{h_n,F_n}\left(\left(g(A_{i^*j^*}) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)|\xi_i^*\right)\right)^2\right) \\
&\overset{LIE}{=} E_{h_n,F_n}\left(\left(E_{h_n,F_n}\left(\left(E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)|\xi_i^*\right)\right)^2\right) = 0.
\end{aligned}
$$

- If there is full overlap ($i^* = k^*$ and $j^* = l^*$, or $i^* = l^*$ and $j^* = k^*$):

$$
E_{h_n,F_n}\left(\left(g(A_{i^*j^*}) - E_{h_n}\left(g(A_{i^*j^*})|\xi^*\right)\right)^2\right) \leq E_{h_n,F_n}\left(g^2(A_{i^*j^*})\right).
$$

  There are $\binom{n}{2}$ terms of this final form in the sum.

Combining the three cases, we get:

$$
E_{h_n,F_n}\left(\left(f_n\left(A^*\left(h_n\left(\xi^*\right),\eta^*\right),\rho_n,F_n\right) - \tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right)\right)^2\right) \quad \leq \frac{2E_{h_n,F_n}\left(g^2(A_{i^*j^*})\right)}{(n-1)\rho_n^{2\tau(g)}}
$$

In an analogous way to how we have defined $\tilde{g}$ and the corresponding $\tilde{\tilde{g}}$, we let[27] $\tilde{g^2}\left(h_{0,n}\left(\xi_i,\xi_j\right)\right) \equiv E_{h_{0,n}}\left(g^2(A_{ij})|\xi\right)$, and we can find a function $\tilde{\tilde{g^2}}(w_0(\xi_i,\xi_j))$ with $\frac{\tilde{g^2}(h_{0,n}(\xi_i,\xi_j))}{\rho_n^{\tau(g)}} = \tilde{\tilde{g^2}}(w_0(\xi_i,\xi_j)) + O(\rho_n)$, $0 < E_{F_0}\left(\left|\tilde{\tilde{g^2}}(w_0(\xi_i,\xi_j))\right|\right) < \infty$ and

[27] Comparing to the example given earlier:

$$
\begin{aligned}
E(g^2(A_{1,2},A_{2,3})|\xi) \equiv \tilde{g^2}(h_{0,n}(\xi_1,\xi_2),h_{0,n}(\xi_2,\xi_3)) \\
&= g^2(0,0)(1-h_{0,n}(\xi_1,\xi_2))(1-h_{0,n}(\xi_2,\xi_3)) + g^2(0,1)(1-h_{0,n}(\xi_1,\xi_2))h_{0,n}(\xi_2,\xi_3) \\
&+ g^2(1,0)h_{0,n}(\xi_1,\xi_2)(1-h_{0,n}(\xi_2,\xi_3)) + g^2(1,1)h_{0,n}(\xi_1,\xi_2)h_{0,n}(\xi_2,\xi_3).
\end{aligned}
$$

This example illustrates why $\tilde{g^2}(h_{0,n}(\xi_1,\xi_2),h_{0,n}(\xi_2,\xi_3))$ is proportional to $\rho_n^{\tau(g)}$, not to $\rho_n^{2\tau(g)}$.

$0 < E_{F_0}\left(\left|\widetilde{\widetilde{g^2}}(w_0(\xi_i,\xi_i))\right|\right) < \infty$. Then:

$$\begin{aligned}
&E_{h_n,F_n}\left(\rho_n^{-\tau(g)} g^2(A_{i^*j^*})\right)\\
&= E_{F_n}\left(E_{h_n}\left(\rho_n^{-\tau(g)} g^2(A_{i^*j^*})|\xi^*\right)\right)\\
&= E_{F_n}\left(\rho_n^{-\tau(g)} \widetilde{g^2}\left(h_n\left(\xi_i^*,\xi_j^*\right)\right)\right)\\
&= E_{F_n}\left(\rho_n^{-\tau(g)} \widetilde{g^2}\left(h_{0,n}\left(\xi_i^*,\xi_j^*\right)\right) + \rho_n^{-\tau(g)+1} \widetilde{g^2}'\left(\tilde{h}_n\left(\xi_i^*,\xi_j^*\right)\right)\frac{1}{\rho_n}\left(h_n\left(\xi_i^*,\xi_j^*\right) - h_{0,n}\left(\xi_i^*,\xi_j^*\right)\right)\right)\\
&\leq E_{F_n}\left(\rho_n^{-\tau(g)} \widetilde{g^2}\left(h_{0,n}\left(\xi_i^*,\xi_j^*\right)\right)\right) + \underbrace{\rho_n^{-\tau(g)+1}\sup_h\left|\widetilde{g^2}'(h)\right|}_{<\infty}\underbrace{E_{F_n}\left(\frac{1}{\rho_n}\left(h_n(\xi_i^*,\xi_j^*) - h_{0,n}(\xi_i^*,\xi_j^*)\right)\right)}_{=o(1)}\\
&= E_{F_n}\left(\widetilde{\widetilde{g^2}}\left(w_0\left(\xi_i^*,\xi_j^*\right)\right)\right) + O\left(\rho_n\right) + o(1)\\
&\xrightarrow{a.s.} E_{F_0}\left(\widetilde{\widetilde{g^2}}(w_0(\xi_i,\xi_j))\right) < \infty.
\end{aligned}$$

Note that since the leading term of $\widetilde{g^2}(h_{0,n})$ is proportional to the $\tau(g)$th power of $h_{0,n}$, the leading term of $\widetilde{g^2}'(h_{0,n})$ has a $h_{0,n}$ to the power $\tau(g)-1$. Given the form of $\widetilde{g^2}'(h_{0,n})$, which is a sum of finitely many terms of the form of a bounded constant times bounded powers of $h_{0,n}$, the whole derivative is bounded. It follows that:

$$E_{h_n,F_n}\left(\left(f_n\left(A^*\left(h_n\left(\xi^*\right),\eta^*\right),\rho_n,F_n\right) - \tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right)\right)^2\right) \leq O\left(\frac{1}{n\rho_n^{\tau(g)}}\right) = o\left(1\right).$$

Hence (iii) holds.

Checking (ii) is a bit more involved. We start with a Hoeffding's (martingale) decomposition[28] of $\tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right)$ for any $\{(h_n,F_n)\}_{n=1}^{\infty}$ in $C_{w_0,F_0,\rho}$:

$$\begin{aligned}
\tilde{f}_n\left(h_n\left(\xi^*\right),\rho_n,F_n\right) &= \frac{\sqrt{n}}{\binom{n}{2}\rho_n^{\tau(g)}}\sum_{i<j}\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))\right)\\
&= \frac{2}{\sqrt{n}\rho_n^{\tau(g)}}\sum_{i=1}^{n} E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))|\xi_i^*\right) - E_{F_n}\left(E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))|\xi_i^*\right)\right)\\
&+ \frac{\sqrt{n}}{\binom{n}{2}\rho_n^{\tau(g)}}\sum_{i<j}\Big(\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))|\xi_i^*\right)\\
&\qquad - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))|\xi_j^*\right) + E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))\right)\Big)\\
&\equiv \tilde{U}_n(h_n,F_n) + \tilde{r}_n(h_n,F_n).
\end{aligned}$$

We firstly focus on $\tilde{U}_n(h_n,F_n)$, which is a (rescaled) average of i.i.d. terms. We add and

[28] For more details see Chapter 5 of Serfling (2009), specifically section 5.1.5.

subtract terms that swap h_n for $h_{0,n}$ and F_n for F_0:

$$
\begin{aligned}
\tilde{U}_n(h_n, F_n) = \frac{2}{\sqrt{n}\rho_n^{\tau(g)}} \sum_{i=1}^{n} \Big(& E_{F_n}\left(\tilde{g}(h_n(\xi_i^*, \xi_j^*))|\xi_i^*\right) - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*, \xi_j^*))\right) \\
& - E_{F_n}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*))|\xi_i^*\right) + E_{F_n}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*))\right) \Big) \\
+ \frac{2}{\sqrt{n}\rho_n^{\tau(g)}} \sum_{i=1}^{n} \Big(& E_{F_n}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*))|\xi_i^*\right) - E_{F_n}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*))\right) \\
& - E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j))|\xi_i^*\right) + E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j))|\xi_i^*\right)\right) \Big) \\
+ \frac{2}{\sqrt{n}\rho_n^{\tau(g)}} \sum_{i=1}^{n} & \left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j))|\xi_i^*\right) - E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*, \xi_j))|\xi_i^*\right)\right)\right) \\
= T_1 + T_2 + T_3 &
\end{aligned}
$$

We deal with these terms one by one.

For T_1, we do Taylor expansion of $\tilde{g}$ around $h_{0,n}$:

$$
\tilde{g}(h_n(\xi_i^*, \xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*)) = \tilde{g}^{'}\left(\tilde{h}_n(\xi_i^*, \xi_j^*)\right)\left(h_n(\xi_i^*, \xi_j^*) - h_{0,n}(\xi_i^*, \xi_j^*)\right)
$$

where $\tilde{h}_n(\xi_i^*, \xi_j^*)$ is between $h_n(\xi_i^*, \xi_j^*)$ and $h_{0,n}(\xi_i^*, \xi_j^*)$. We can show that T_1 goes to zero in second mean, hence also in probability. Let:

$$
\begin{aligned}
T_1 &= \frac{2}{\sqrt{n}\rho_n^{\tau(g)}} \sum_{i=1}^{n} \Big(E_{F_n}\left(\tilde{g}(h_n(\xi_i^*, \xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*))|\xi_i^*\right) \\
&\qquad - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*, \xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*, \xi_j^*))\right) \Big) \\
&= \frac{2}{\sqrt{n}\rho_n^{\tau(g)}} \sum_{i=1}^{n} b_{i^*}.
\end{aligned}
$$

Note that the terms inside the sum are independent and have zero expectation:

$$
\begin{aligned}
E_{F_n}(b_{i^*}) = E_{F_n}\Big(& E_{F_n}\left(\tilde{g}(h_n(\xi_i^*, \xi_k^*)) - \tilde{g}(h_{0,n}(\xi_i^*, \xi_k^*))|\xi_i^*\right) \\
& - E_{F_n}\left(\tilde{g}(h_n(\xi_k^*, \xi_l^*)) - \tilde{g}(h_{0,n}(\xi_k^*, \xi_l^*))\right) \Big) \\
\overset{LIE}{=} & \; E_{F_n}\left(\tilde{g}(h_n(\xi_k^*, \xi_l^*)) - \tilde{g}(h_{0,n}(\xi_k^*, \xi_l^*))\right) - E_{F_n}\left(\tilde{g}(h_n(\xi_k^*, \xi_l^*)) - \tilde{g}(h_{0,n}(\xi_k^*, \xi_l^*))\right) = 0
\end{aligned}
$$

Hence in the expansion of the square all terms with $i \neq j$ are zero:

$$
\begin{aligned}
E\left(T_1^2\right) &= E\left(\left(\frac{2}{\sqrt{n}\rho_n^{\tau(g)}}\sum_{i=1}^{n} b_{i^*}\right)^2\right) \\
&= 4\frac{1}{n\rho_n^{2\tau(g)}}\sum_{i=1}^{n} E_{F_n}\left(b_{i^*}^2\right) + 8\frac{1}{n\rho_n^{2\tau(g)}}\sum_{i<j} E_{F_n}\left(b_{i^*}b_{j^*}\right) \\
&\overset{i.i.d.}{=} 4\rho_n^{-2\tau(g)}E_{F_n}\left(b_{i^*}^2\right) + 8\frac{1}{n\rho_n^{2\tau(g)}}\sum_{i<j} E_{F_n}\left(b_{i^*}\right)E_{F_n}\left(b_{j^*}\right) \\
&= 4\rho_n^{-2\tau(g)}E_{F_n}\left(b_{i^*}^2\right) \\
&= 4\rho_n^{-2\tau(g)}E_{F_n}\Bigg(\bigg(E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*,\xi_j^*))|\xi_i^*\right) \\
&\qquad\qquad - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*,\xi_j^*))\right)\bigg)^2\Bigg) \\
&= 4\rho_n^{-2\tau(g)}E_{F_n}\left(\left(E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*,\xi_j^*))|\xi_i^*\right)\right)^2\right) \\
&\quad - 4\rho_n^{-2\tau(g)}\left(E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - \tilde{g}(h_{0,n}(\xi_i^*,\xi_j^*))\right)\right)^2 \\
&\leq 4\rho_n^{-2\tau(g)}E_{F_n}\left(\left(E_{F_n}\left(\tilde{g}^{'}\left(\tilde{h}_n(\xi_i^*,\xi_j^*)\right)\left(h_n(\xi_i^*,\xi_j^*) - h_{0,n}(\xi_i^*,\xi_j^*)\right)|\xi_i\right)\right)^2\right) \\
&\leq 4\underbrace{\left(\rho_n^{-\tau(g)+1}\sup_{h\in[0,M_w\rho_n]}\left|\tilde{g}^{'}(h)\right|\right)^2}_{<\infty}\underbrace{E_{F_n}\left(\left(\frac{1}{\rho_n}\left(h_n(\xi_i^*,\xi_j^*) - h_{0,n}(\xi_i^*,\xi_j^*)\right)\right)^2\right)}_{=o(1)} \to 0
\end{aligned}
$$

In the first inequality we use the fact that the second term is negative and smaller in magnitude than the first. We then pull the supremum over derivatives of $\tilde{g}$ out of the expectation, use Jensen's inequality to put the square inside the inner expectation, apply the law of iterated expectations, and use the assumption 1. to get the conclusion.

As mentioned before, the derivative of $\tilde{g}$ is bounded for any choice of g, and as we take a derivative with respect to h the leading term of the $\tilde{g}^{'}$ becomes proportional to power one lower than $\tilde{g}$, i.e. $\rho_n^{-\tau(g)+1}\tilde{g}^{'} = O_p(1)$.

For the middle term, T_2, we show that it goes to zero in mean squared. To simplify notation, let $T_2 = \frac{2}{\sqrt{n}\rho_n^{\tau(g)}}\sum_{i=1}^{n} a_{i^*}$ and notice that:

$$
\begin{aligned}
E_{F_n}\left(a_{i^*}\right) = &\underbrace{E_{F_n}\left(E_{F_n}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j^*))|\xi_i^*\right)\right) - E_{F_n}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j^*))\right)}_{=0} \\
&\underbrace{-E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right)\right) + E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right)\right)}_{=0} = 0.
\end{aligned}
$$

Then we have:

$$\begin{aligned}
E\left(T_2^2\right) &= E\left(\left(\frac{2}{\sqrt{n}\rho_n^{\tau(g)}}\sum_{i=1}^{n} a_{i^*}\right)^2\right) \\
&= 4\frac{1}{n\rho_n^{2\tau(g)}}\sum_{i=1}^{n} E_{F_n}\left(a_{i^*}^2\right) + 8\frac{1}{n\rho_n^{2\tau(g)}}\sum_{i<j} E_{F_n}\left(a_{i^*}a_{j^*}\right) \\
&\overset{i.i.d.}{=} 4\rho_n^{-2\tau(g)}E_{F_n}\left(a_{i^*}^2\right) + 8\frac{1}{n\rho_n^{2\tau(g)}}\sum_{i<j} E_{F_n}\left(a_{i^*}\right)E_{F_n}\left(a_{j^*}\right) \\
&= 4\rho_n^{-2\tau(g)}E_{F_n}\left(a_{i^*}^2\right) \\
&= 4E_{F_n}\Big(\big(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right) + E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))\right) - E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right) \\
&\quad + E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\right) - E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\right) + E_{F_n}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\big)^2\Big) \\
&\quad + O\left(\rho_n\right) \\
&\leq 8\underbrace{E_{F_n}\left(\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)\right)^2\right)}_{\to E_{F_0}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)^2\right)} + 8\underbrace{E_{F_n}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)^2\right)}_{\to E_{F_0}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)^2\right)} \\
&\quad - 16\underbrace{E_{F_n}\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)}_{\to E_{F_0}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)^2\right)} \\
&\quad + 8\underbrace{\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))\right) - E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\right)\right)^2}_{\to 0} \\
&\quad + 8\underbrace{\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\right) - E_{F_n}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\right)^2}_{\to 0} + O\left(\rho_n\right) \\
&\to 0
\end{aligned}$$

In the 5th equality we plug in the definition of a_{i^*}, we add and subtract the term $E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\right)$, we bring the normalisation by $\rho_n^{-2\tau(g)}$ inside the expectation and use $\frac{\tilde{g}(\xi_i,\xi_j)}{\rho_n^{\tau(g)}} = \tilde{\tilde{g}}\left(w_0\left(\xi_i,\xi_j\right)\right) + O\left(\rho_n\right)$. In the next step, we apply $(a+b)^2 \leq 2a^2 + 2b^2$, where a corresponds to the first four terms in the previous summation, for which we expand the square, and b corresponds to the last two terms. We now verify that we can apply property 2. to all resulting terms:

- By the independence between ξ_j^* and ξ_k^* when $j \neq k$ and the law of iterated expectations we can rewrite the first term as:

$$\begin{aligned}
E_{F_n}\left(\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)\right)^2\right) &= E_{F_n}\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_k^*))|\xi_i^*\right)\right) \\
&= E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_k^*))\right)
\end{aligned}$$

We now check the conditions for 2. when all indices are unique:

$$\begin{aligned}
E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\right) &\overset{LIE}{=} E_{F_0}\left(E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\,|\,\xi_i\right)\right)\\
&\leq E_{F_0}\left(E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\,|\,\xi_i\right)E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\,|\,\xi_i\right)\right)\\
&= E_{F_0}\left(E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\,|\,\xi_i\right)^2\right)\\
&\leq E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\big|\,\xi_i\right)\right)\\
&\overset{LIE}{=} E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\right) < \infty.
\end{aligned}$$

When two indices are repeated we use Cauchy-Schwarz inequality:

$$E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))|\right) \leq \sqrt{E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\right)E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_i))\right)} < \infty.$$

And when all indices are equal the condition $E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_i))\right) < \infty$ follows straight from the assumptions. Hence we have

$$\begin{aligned}
E_{F_n}\left(\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)\right)^2\right) &\to E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\right)\\
&= E_{F_0}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)^2\right).
\end{aligned}$$

- For the second term, we can verify the condition for 2. when the indices are unique:

$$\begin{aligned}
E_{F_0}\left(\left|E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)^2\right|\right) &= E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)^2\right)\\
&\leq E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))|\xi_i\right)\right)\\
&\overset{LIE}{=} E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\right) < \infty,
\end{aligned}$$

where the inequality follows from Jensen's inequality. When the indices are repeated:

$$E_{F_0}\left(\left|E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))|\xi_i\right)^2\right|\right) = E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_i))\right) < \infty.$$

hence $E_{F_n}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)^2\right) \to E_{F_0}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)^2\right)$.

- The third term can be rewritten as:

$$\begin{aligned}
&E_{F_n}\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\\
&\quad= E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right).
\end{aligned}$$

Using Jensen's inequality, we verify the condition for 2. when the indices are unique:

$$\begin{aligned}
&E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\xi_i\right)|\right)\\
&\quad\le E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\, E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\,|\xi_i\right)\right)\\
&\quad\overset{LIE}{=} E_{F_0}\left(E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\,|\xi_i\right) E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_k))|\,|\xi_i\right)\right)\\
&\quad= E_{F_0}\left(E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\,|\xi_i\right)^2\right)\\
&\quad\le E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))|\xi_i\right)\right)\\
&\quad\overset{LIE}{=} E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\right)<\infty
\end{aligned}$$

and using Jensen's and Cauchy-Schwarz inequalities we verify it when the indices are equal:

$$\begin{aligned}
E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)|\right) &\le E_{F_0}\left(E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\,|\xi_i\right)\right)\\
&\overset{LIE}{=} E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\right)\\
&\le \sqrt{E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\right)E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_i))\right)}\\
&<\infty.
\end{aligned}$$

hence

$$\begin{aligned}
&E_{F_n}\left(E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))|\xi_i^*\right)E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\\
&\qquad\to E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)=E_{F_0}\left(\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)^2\right)
\end{aligned}$$

- For the fourth term we can verify that

$$\begin{aligned}
E_{F_0}\left(|E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)|\right) &\le \sqrt{E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))|\xi_i\right)\right)}\\
&\overset{LIE}{=} \sqrt{E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_j))\right)}<\infty,\\
E_{F_0}\left(|E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))|\xi_i\right)|\right) = E_{F_0}\left(|\tilde{\tilde{g}}(w_0(\xi_i,\xi_i))|\right) &\le \sqrt{E_{F_0}\left(\tilde{\tilde{g}}^2(w_0(\xi_i,\xi_i))\right)}<\infty.
\end{aligned}$$

hence $E_{F_n}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j^*))\right)\to E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))\right)$.

We combine all terms using continuous mapping theorem and see that they all cancel out and the limit is zero.

For T_3, we can write:

$$\begin{aligned}T_3 = 2\rho_n^{-\tau(g)}&\sqrt{Var_{F_n}(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right))}\\ &\times\sum_{i=1}^n \frac{E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right) - E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right)\right)}{\sqrt{n}\sqrt{Var_{F_n}(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right))}}.\end{aligned}$$

Denote the terms inside the sum by X_{in}. They have zero expectation:

$$E_{F_n}\left(X_{in}\right) = \frac{E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right) - E_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\right)}{\sqrt{n}\sqrt{Var_{F_n}(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right))}} = 0.$$

Their variances sum to 1 for each n:

$$\sum_{i=1}^n Var_{F_n}\left(X_{in}\right) = n\frac{Var_{F_n}(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right))}{nVar_{F_n}(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right))} = 1.$$

And for all n when $i \neq j$ the terms X_{in} and X_{jn} are independent and identically distributed. Hence by Lindeberg-Levy CLT for triangular arrays their sum converges in distribution to a standard normal random variable.

For the multiplier term we have:

$$\begin{aligned}\rho_n^{-2\tau(g)}Var_{F_n}\left(E_{F_0}\left(\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right)\right) &= Var_{F_n}\left(E_{F_0}\left(\rho_n^{-\tau(g)}\tilde{g}(h_{0,n}(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\\ &= Var_{F_n}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right) + O(\rho_n)\\ &= E_{F_n}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)^2\right) - \left(E_{F_n}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i^*,\xi_j))|\xi_i^*\right)\right)\right)^2 + O(\rho_n)\\ &\to E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)^2\right) - \left(E_{F_0}\left(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)\right)\right)^2\\ &= Var_{F_0}(E_{F_0}\left(\tilde{\tilde{g}}(w_0(\xi_i,\xi_j))|\xi_i\right)) \equiv \sigma_1^2 < \infty.\end{aligned}$$

The first equality is pulling the normalisation inside the variance. The second equality applies the definition og $\tilde{\tilde{g}}$. The third equality is rewriting variance in terms of expectations. The limit follows from 2. (we have already checked that the relevant absolute moments are finite when we were checking conditions for convergence of T_2, terms two and four) and the continuous mapping theorem. The final line is by definition. Hence $T_3 \xrightarrow{d} N(0, 4\sigma_1^2)$.

It remains to show that $\tilde{r}_n(h_n, F_n) = o_p(1)$. We can check that the expression is $\sqrt{n}$ times a U-statistic with a kernel function $G(\xi_i^*,\xi_j^*) = \rho_n^{-\tau(g)}\big(\tilde{g}(h_n(\xi_i^*,\xi_j^*)) - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))|\xi_i^*\right) - E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))|\xi_j^*\right) + E_{F_n}\left(\tilde{g}(h_n(\xi_i^*,\xi_j^*))\right)\big)$. Note that $E(G(\xi_i^*,\xi_j^*)) = E(G(\xi_i^*,\xi_j^*)|\xi_i^*) = 0$, i.e. it is a degenerate U-statistic with $Var(E(G(\xi_i^*,\xi_j^*)|\xi_i^*)) = 0$. We could show that the whole term is negligible by convergence in second mean from definition, or rely on a Theorem from section 5.3.2 in Serfling (2009) which, in the present setting, can be stated as:

Lemma (Theorem 5.3.2 in Serfling (2009)). *If* $E_{F_n}\left(\left(G(\xi_i^*,\xi_j^*)\right)^2\right) < \infty$ *then*

$$E_{F_n}\left(\left(\tilde r_n(h_n,F_n)\right)^2\right) = O\left(\frac{1}{n}\right) = o(1).$$

By Jensen's inequality and the law of large numbers, $\rho_n^{-2\tau(g)} E_{F_n}\left(\tilde g^2\left(h_n\left(\xi_i^*,\xi_j^*\right)\right)\right)$ is an upper bound for all terms in the expansion of $E_{F_n}\left(\left(G(\xi_i^*,\xi_j^*)\right)^2\right)$. Hence the sufficient condition is implied by:

$$\begin{aligned}
\rho_n^{-2\tau(g)} E_{F_n}\left(\tilde g^2\left(h_n\left(\xi_i^*,\xi_j^*\right)\right)\right) &\le 2\underbrace{E_{F_n}\left(\tilde{\tilde g}^2\left(w_0\left(\xi_i^*,\xi_j^*\right)\right)\right)}_{\to E_{F_0}\left(\tilde{\tilde g}^2(w_0(\xi_i,\xi_j))\right)<\infty} + O(\rho_n) \\
&\quad + 2\underbrace{\left(\sup_{h\in[0,M_w\rho_n]}\left|\frac{\tilde g^{'}(h)}{\rho_n^{\tau(g)-1}}\right|\right)^2}_{<\infty}\underbrace{E_{F_n}\left(\left(\frac{1}{\rho_n}\left(h_n\left(\xi_i^*,\xi_j^*\right)-\hat h_n\left(\xi_i^*,\xi_j^*\right)\right)\right)^2\right)}_{\to 0} \\
&\le 2E_{F_0}\left(\tilde{\tilde g}^2\left(w_0\left(\xi_i,\xi_j\right)\right)\right) + o(1).
\end{aligned}$$

Hence for any $\varepsilon > 0$ we can find an N sufficiently large so that the condition is satisfied: $E_{F_n}\left(\left(G(\xi_i^*,\xi_j^*)\right)^2\right) < 8E_{F_0}\left(\tilde{\tilde g}^2\left(w_0\left(\xi_i,\xi_j\right)\right)\right) + \varepsilon < \infty$ for all $n > N$.

Moving on to the second part of the proof, we check that the sequence $\left\{\hat h_n, \hat F_n\right\}_{n=1}^{\infty}$ satisfies assumptions 1. and 2. in probability:

1. Follows from Theorem 1:

$$\begin{aligned}
&E_{\hat F_n}\left(\left(\frac{1}{\rho_n}\left(\hat h_n(\xi_i^*,\xi_j^*) - h_{0,n}(\xi_i^*,\xi_j^*)\right)\right)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{\rho_n}\left(\hat h_n(\xi_i,\xi_j) - h_{0,n}(\xi_i,\xi_j)\right)\right)^2 \\
&\le \left(\max_{i,j}\left|\frac{\hat h_n(\xi_i,\xi_j) - h_{0,n}(\xi_i,\xi_j)}{\rho_n}\right|\right)^2 \\
&= o_p(1)^2 = o_p(1).
\end{aligned}$$

2. Let $f : Supp(\xi)^3 \to \mathbb{R}$ be any symmetric function for which $E_{F_0}\left(\left|f\left(\xi_i,\xi_j,\xi_k\right)\right|\right) < \infty$,

$E_{F_0}\left(\left|f\left(\xi_i,\xi_i,\xi_j\right)\right|\right)<\infty$ and $E_{F_0}\left(\left|f\left(\xi_i,\xi_i,\xi_i\right)\right|\right)<\infty$. We have:

$$
\begin{aligned}
&E_{\hat{F}_n}\left(f(\xi_i^*,\xi_j^*,\xi_k^*)\right)=\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}f(\xi_i,\xi_j,\xi_k)\\
&=\underbrace{\frac{(n-1)(n-2)}{n^2}}_{\to 1}\underbrace{\frac{1}{\binom{n}{3}}\sum_{i<j<k}f(\xi_i,\xi_j,\xi_k)}_{\xrightarrow{a.s.}E_{F_0}(f(\xi_i,\xi_j,\xi_k))<\infty}+\underbrace{\frac{3(n-1)}{n^2}}_{\to 0}\underbrace{\frac{1}{\binom{n}{2}}\sum_{i<j}f(\xi_i,\xi_i,\xi_j)}_{\xrightarrow{a.s.}E_{F_0}(f(\xi_i,\xi_i,\xi_j))<\infty}\\
&\quad+\underbrace{\frac{1}{n^2}}_{\to 0}\underbrace{\frac{1}{n}\sum_{i=1}^{n}f(\xi_i,\xi_i,\xi_i)}_{\xrightarrow{a.s}E_{F_0}(f(\xi_i,\xi_i,\xi_i))<\infty}\\
&\xrightarrow{a.s.}E_{F_0}\left(f(\xi_i,\xi_j,\xi_k)\right)
\end{aligned}
$$

The first equality follows from the definition of the empirical distribution function $\hat{F}_n$. The convergence of the two terms in the second line follows from the SLLN for U-statistics (see e.g. Theorem A. in section 5.4 of Serfling (2009), p.190) given that $E_{F_0}\left(\left|f\left(\xi_i,\xi_j,\xi_k\right)\right|\right)<\infty$ and $E_{F_0}\left(\left|f\left(\xi_i,\xi_i,\xi_j\right)\right|\right)<\infty$. The convergence of the term in the third line follows from Kolmogorov's SLLN for i.i.d. random variables which applies under the assumption that $E_{F_0}\left(\left|f\left(\xi_i,\xi_i,\xi_i\right)\right|\right)<\infty$. The final line is by continuous mapping theorem for almost sure convergence. Condition 2. holds almost surely (hence also in probability), but because we only get condition 1. in probability the overall result is for convergence weakly in probability.

□

The above result was stated for a normalisation using the unknown ρ_n. We now show that the conclusions remain true when we replace it with an estimate.

Corollary 2.A.1. *Under the assumptions of Theorem 2.A.2*

$$\hat{\rho}_n-\rho_n=o_p(\rho_n),$$

hence

$$f_n\left(A\left(\hat{h}_n(\xi^*),\eta^*\right),\hat{\rho}_n,\hat{F}_n\right)=f_n\left(A\left(\hat{h}_n(\xi^*),\eta^*\right),\rho_n,\hat{F}_n\right)+o_p(1)$$

and we get all conclusions of Theorem 2.A.1 in probability with $J(t,w_0,F_0)=N(0,m^2\sigma_1^2)$ for

$$
\begin{aligned}
f_n\left(A\left(\hat{h}_n\left(\xi^*\right),\eta^*\right),\hat{\rho}_n,\hat{F}_n\right)=&\frac{\sqrt{n}}{\binom{n}{m}\hat{\rho}_n^{\tau(g)}}\sum_{1\leq\iota_1<\cdots<\iota_m\leq n}\left(g\left(A\left(\hat{h}_n\left(\xi^*\left(\iota\right)\right),\eta^*\left(\iota\right)\right)\right)\right.\\
&\left.-E_{\hat{h}_n,\hat{F}_n}\left(g\left(A\left(\hat{h}_n\left(\xi^*\left(\iota\right)\right),\eta^*\left(\iota\right)\right)\right)\right)\right).
\end{aligned}
$$

Proof of Corollary 2.A.1. Let ρ_n^* denote the density of a bootstrap adjacency matrix formed by $\xi^* \sim F_n$ with linking probabilities h_n. We can write

$$f_n\left(A\left(h_n\left(\xi^*\right), \eta^*\right), \rho_n^*, F_n\right) = \left(\frac{\rho_n}{\hat{\rho}_n^*}\right)^{\tau(g)} f_n\left(A\left(h_n\left(\xi^*\right), \eta^*\right), \rho_n, F_n\right)$$

hence it is sufficient to show $\rho_n^* - \rho_n = o_p(\rho_n)$ which, by Slutsky's theorem, implies $\left(\frac{\rho_n}{\rho_n^*}\right)^{\tau(g)} \xrightarrow{p} 1$.

Applying Theorem 2.4.2 to $g(A_{ij}) = A_{ij}$, for which $m = 2$ and $\tau(g) = 1$, we have:

$$\frac{\sqrt{n}}{\binom{n}{2}\rho_n} \sum_{1 \leq i^* < j^* \leq n} \left(A_{i^*j^*} - E_{h_n,F_n}(A_{i^*j^*})\right) = O_p(1)$$

where the expression is bounded in probability because it has a well-defined limiting distribution.

Hence

$$\begin{aligned}
\rho_n^* &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i^* < j^* \leq n} A_{i^*j^*} \\
&= E_{h_n,F_n}(A_{i^*j^*}) + \underbrace{\frac{\rho_n}{\sqrt{n}} O_p(1)}_{=\frac{1}{\sqrt{n}} O_p(\rho_n) = o_p(\rho_n)} \\
&= E_{F_n}\left(E_{h_n}\left(A_{i^*j^*} \middle| \xi^*\right)\right) + o_p(\rho_n) \\
&= E_{F_n}\left(E_{h_n}\left(h_n(\xi_i^*, \xi_j^*) \middle| \xi^*\right)\right) + o_p(\rho_n) \\
&= \rho_n\left(E_{F_n}\left(\frac{1}{\rho_n}\left(h_n(\xi_i^*, \xi_j^*) - h_{0,n}(\xi_i^*, \xi_j^*)\right)\right) + E_{F_n}\left(\frac{1}{\rho_n} h_{0,n}(\xi_i^*, \xi_j^*)\right)\right) + o_p(\rho_n) \\
&\leq \rho_n\left(\sqrt{E_{F_n}\left(\left(\frac{1}{\rho_n}\left(h_n(\xi_i^*, \xi_j^*) - h_{0,n}(\xi_i^*, \xi_j^*)\right)\right)^2\right)} + E_{F_n}\left(w_0(\xi_i^*, \xi_j^*)\right)\right) + o_p(\rho_n) \\
&= \rho_n\left(o(1) + E_{F_0}\left(w_0\left(\xi_i, \xi_j\right) + o(1)\right)\right) + o_p(\rho_n) \\
&= \rho_n(o(1) + 1 + o(1)) + o_p(\rho_n) \\
&= \rho_n + o_p(\rho_n)
\end{aligned}$$

The first equality is by definition, the second follows from the above expression and result from Theorem 2.4.2. The third equality uses the law of iterated expectations. The fourth equality is by definition of A. For the fifth equality we add and subtract $E_{F_n}\left(h_{0,n}(\xi_i^*, \xi_j^*)\right)$ and pull ρ_n out of the bracket. The inequality is due to Jensen's inequality where the final term is transformed according to the definition of w_0. The sixth equality uses assumptions 1. and 2.. The seventh equality is due to the definition of w_0 which is assumed to integrate to 1.

The above derivation applies to the case of $\rho_n^* = \hat{\rho}_n$ (for $h_n = h_{0,n}$ and $f_n = F_0$), proving that $f_n\left(A\left(h_{0,n}\left(\xi\right),\eta\right),\hat{\rho}_n,F_0\right)$ and $f_n\left(A\left(h_{0,n}\left(\xi\right),\eta\right),\rho_n,F_0\right)$ have the same asymptotic limit.

If we replaces ρ_n^* with $\hat{\rho}_n^*$ the $o(1)$ terms in the derivation are replaced by $o_p(1)$, which does not affect the overall result. Hence we also get the same limit of $f_n\left(A\left(\hat{h}_n\left(\xi^*\right),\eta^*\right),\hat{\rho}_n^*,\hat{F}_n\right)$.

□

We note that all conclusions of Theorem 2.4.2, Lemma 2.4.1 and Corollary 2.4.1 follow from Theorem 2.A.2 and Corollary 2.A.1, hence they have also been proven.

Appendix 2.B Additional tables, plots, codes

Subsection 2.B.1 Codes

In this section we present some of the codes used for simulations. A full package should eventually become available online.

Note that these codes are used together with those given in Chapter 1 in Section 1.B.2.

Code for running bootstrap:

```
def boot_HK1h(A,h,B):
#outputs B bootstrapped adjacency matrices based on matrix A with bandwidth h using linking function estimate HK1
n=len(A)
H_true = HK1h(D2(A),A,h)
#choose nodes for bootstrap villages:
v = np.random.randint(0, n, size=(B,n))
#generate new adjacency matrices
row = np.tensordot(v,np.ones(n),0).astype(int)
column = np.tile(np.array(v),n).reshape(B,n,n)
G = H_true[row,column]
u = np.random.rand(B, n, n)
m = np.tril(u) + np.transpose(np.tril(u, -1),[0,2,1])
A_boot = (m < G)*1
[np.fill_diagonal(A_boot[i], 0) for i in range(B)]
return A_boot
```

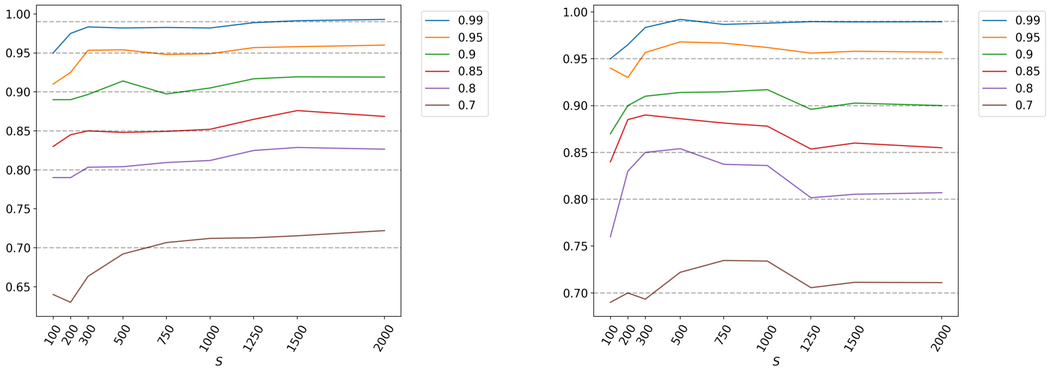
A sample Monte Carlo simulation code using the above definitions and the WARP procedure from Giacomini, Politis, and White (2013) to obtain the confidence interval coverage:

```
#define the output data frame:
df_loo = pd.DataFrame(columns=['S', 'B', 'n', 'rho','average for true graphs', 'true value', 'alpha', 'proportion of bootstrap CI that cover truth', 'average length of bootstrap CI', 'statistic','h'])
```

```
#run the simulations:
ALPHA = [0.01, 0.05, 0.1, 0.15, 0.2, 0.3] #sizes of confidence intervals
NN = [25, 50, 100, 150, 200, 300] #sample sizes
SS = [1000] #number of true graphs
RR = [1, 0.75,0.5,0.25,0.1] #sparsity level

for n in NN:
    for r in RR:
        S_max = max(SS)
        (A_true, h_true, xi_true) = product_generate_A_h_xi(n, r, S_max)
        A_boot = []
        h_list = []
        B=1 #because of WARP we only need one bootstrap replication

        true_density = r*0.25

        for A in A_true[0:min(len(A_true),S_max)]:
            #find the optimal bandwidth by minimising the leave-one-out log-likelihood
            h_guess = 0.2090189845643738*true_density**1.38258532*n**(-1.55268817)*np.log(n)**1.82661653
            res = minimize(ll, h_guess, args=A, method = 'Nelder-Mead', tol=1e-7, bounds=((0,1.1),))
            h = res.x[0]
            #do bootstrap for matrix A using the optimal bandwidth
            Ab = boot_HK1h(A,h,B)
            #save the bootstrapped adjacency matrices and the bandwidth
            A_boot.append(Ab)
            h_list.append(h)

        #estimate the statistic of interest for true and bootstrapped graphs
        true_density_all = [nx.density(nx.from_numpy_array(A_true[s])) for s in range(S_max)]]
        true_density_mean = np.mean(true_density_all)
        boot_density_all = [nx.density(nx.from_numpy_array(A_boot[s][0])) for s in range(S_max)]

        #find the confidence interval coverage using WARP
        for S in SS:
            if (S<=len(A_true)):
                true_density_vec = true_density_all[0:S]
                boot_density = boot_density_all[0:S]
                stat = 'density'
                density_minus_true = np.array(boot_density) - np.array(
                                                        true_density_vec)
                for alpha in ALPHA:
                    qu= np.percentile(density_minus_true, 100*(1-alpha/2))
                    ql= np.percentile(density_minus_true, 100*(alpha/2))
                    bl = true_density_vec - qu
                    bu = true_density_vec - ql
                    co= np.mean([bl[i] <= true_density <= bu[i] for i in range(S
                                                            )])
                    me = np.mean(bu-bl)
                    df_loo = df_loo._append({'S': S, "B": B, 'n': n, "rho": r, '
                                                            average for true
                                                            graphs':
                                                            true_density_mean, '
                                                            true value':
                                                            true_density, 'alpha
                                                            ': alpha, '
                                                            proportion of
                                                            bootstrap CI that
                                                            cover truth': co, '
                                                            average length of
                                                            bootstrap CI': me, '
                                                            statistic': stat, 'h
                                                            ': np.mean(h_list)},
                                                             ignore_index = True
                                                            )

        #save the output
        df_loo.to_csv('df_loo_product_n_25_300_true_dens.csv')
```

Subsection 2.B.2 Monte Carlo simulations: tables and a sensitivity check

Since we are using the WARP procedure instead of traditional Monte Carlo simulations, we test its sensitivity by checking the effect of varying the number of simulated true graphs S rather than the number of bootstrap replications B, which is always kept at $B = 1$. Fig. 2.11a shows that the predictions for different statistics stabilise above S around 750 or higher. This is true in most simulations (see Table 2.4, Table 2.5), with the exception of networks with high density such as that in Fig. 2.11b (and Table 2.6) in which the predictions do not stabilise until $S = 1250$ or even $S = 1500$. In all other sections we use $S = 1000$. Running more repetitions is computationally expensive and provides little advantage in terms of accuracy in the majority of cases.



(a) Confidence interval coverage for density using the product generating function at $n = 500$ and $\rho_n = 0.1875$.

(b) Confidence interval coverage for transitivity using the high density generating function at $n = 500$ and $\rho_n = 0.759$.

Figure 2.11: Confidence interval coverage for different number of simulated true graphs S based on Monte Carlo simulations.

			average for	Proportion of bootstrap CI that cover truth for					
n	statistic	ρ_n	true graphs	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
500	λ_1	0.024	17.116	1.000	0.971	0.925	0.858	0.804	0.660
		0.078	52.502	0.998	0.961	0.900	0.854	0.811	0.706
		0.139	93.282	0.997	0.956	0.884	0.844	0.799	0.685
		0.250	166.453	0.995	0.954	0.887	0.825	0.778	0.692
	λ_3	0.024	7.823	0.005	0.000	0.000	0.000	0.000	0.000
		0.078	12.825	0.000	0.000	0.000	0.000	0.000	0.000
		0.139	15.582	0.000	0.000	0.000	0.000	0.000	0.000
		0.250	17.042	0.000	0.000	0.000	0.000	0.000	0.000
	λ_{10}	0.024	7.022	0.997	0.985	0.943	0.876	0.803	0.596
		0.078	11.627	0.391	0.058	0.018	0.008	0.004	0.002
		0.139	14.197	0.082	0.000	0.000	0.000	0.000	0.000
		0.250	15.667	0.000	0.000	0.000	0.000	0.000	0.000
	Louvain CDA modularity	0.024	0.243	0.000	0.000	0.000	0.000	0.000	0.000
		0.078	0.119	0.001	0.000	0.000	0.000	0.000	0.000
		0.139	0.082	0.015	0.001	0.001	0.001	0.000	0.000
		0.250	0.052	0.198	0.061	0.021	0.008	0.006	0.002
	density	0.024	0.024	0.405	0.177	0.096	0.059	0.044	0.024
		0.078	0.078	0.989	0.909	0.811	0.752	0.699	0.620
		0.139	0.139	0.993	0.955	0.889	0.827	0.771	0.672
		0.250	0.250	0.992	0.953	0.897	0.839	0.788	0.659
	max betweenness centrality	0.024	0.019	0.304	0.058	0.033	0.023	0.013	0.010
		0.078	0.011	1.000	0.965	0.809	0.692	0.591	0.439
		0.139	0.009	1.000	0.995	0.981	0.923	0.847	0.668
		0.250	0.008	1.000	0.997	0.969	0.901	0.835	0.713
	transitivity	0.024	0.043	1.000	1.000	1.000	0.997	0.993	0.952
		0.078	0.138	1.000	0.967	0.928	0.899	0.864	0.774
		0.139	0.247	0.985	0.954	0.903	0.859	0.818	0.719
		0.250	0.443	0.991	0.947	0.890	0.835	0.779	0.672
	triangle density	0.024	0.000	0.990	0.852	0.706	0.563	0.456	0.309
		0.078	0.001	0.998	0.965	0.905	0.845	0.798	0.701
		0.139	0.006	0.991	0.959	0.902	0.840	0.794	0.692
		0.250	0.037	0.992	0.956	0.882	0.828	0.779	0.678
1000	λ_1	0.013	18.974	0.894	0.677	0.470	0.379	0.270	0.157
		0.058	77.998	0.991	0.951	0.886	0.847	0.812	0.715
		0.120	160.817	0.988	0.959	0.911	0.859	0.814	0.749
		0.250	332.967	0.987	0.959	0.904	0.852	0.810	0.747
	λ_3	0.013	8.573	0.000	0.000	0.000	0.000	0.000	0.000
		0.058	16.482	0.000	0.000	0.000	0.000	0.000	0.000
		0.120	21.588	0.000	0.000	0.000	0.000	0.000	0.000
		0.250	24.642	0.000	0.000	0.000	0.000	0.000	0.000
	λ_{10}	0.013	8.002	0.044	0.000	0.000	0.000	0.000	0.000
		0.058	15.493	0.000	0.000	0.000	0.000	0.000	0.000
		0.120	20.377	0.000	0.000	0.000	0.000	0.000	0.000
		0.250	23.414	0.000	0.000	0.000	0.000	0.000	0.000
	Louvain CDA modularity	0.013	0.234	0.000	0.000	0.000	0.000	0.000	0.000
		0.058	0.100	0.000	0.000	0.000	0.000	0.000	0.000
		0.120	0.064	0.000	0.000	0.000	0.000	0.000	0.000
		0.250	0.037	0.000	0.000	0.000	0.000	0.000	0.000
	density	0.013	0.013	0.000	0.000	0.000	0.000	0.000	0.000
		0.058	0.058	0.937	0.849	0.777	0.705	0.669	0.576
		0.120	0.120	0.984	0.943	0.892	0.836	0.800	0.725
		0.250	0.250	0.992	0.953	0.896	0.846	0.800	0.732
	max betweenness centrality	0.013	0.011	0.101	0.010	0.006	0.004	0.003	0.001
		0.058	0.006	1.000	0.960	0.872	0.776	0.690	0.531
		0.120	0.005	0.998	0.994	0.975	0.957	0.916	0.775
		0.250	0.004	1.000	0.993	0.982	0.947	0.899	0.791
	transitivity	0.013	0.024	1.000	1.000	0.998	0.982	0.955	0.880
		0.058	0.103	0.997	0.980	0.942	0.906	0.864	0.778
		0.120	0.214	0.983	0.952	0.913	0.868	0.835	0.733
		0.250	0.444	0.988	0.959	0.922	0.880	0.827	0.716
	triangle density	0.013	0.000	0.385	0.144	0.064	0.040	0.027	0.011
		0.058	0.000	0.991	0.946	0.884	0.846	0.804	0.714
		0.120	0.004	0.988	0.967	0.921	0.857	0.809	0.745
		0.250	0.037	0.984	0.966	0.910	0.856	0.807	0.746

Table 2.1: Confidence interval coverage for different densities based on Monte Carlo simulations using the product generating function when $S = 1000$.

			average for	Proportion of bootstrap CI that cover truth for					
n	statistic	ρ_n	true graphs	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
250	λ_1	0.127	33.515	0.993	0.961	0.885	0.805	0.744	0.614
		0.307	79.477	0.995	0.965	0.932	0.892	0.851	0.754
		0.476	122.926	0.992	0.976	0.938	0.901	0.845	0.748
		0.740	190.475	0.989	0.970	0.932	0.886	0.838	0.732
	λ_3	0.127	10.044	0.000	0.000	0.000	0.000	0.000	0.000
		0.307	13.573	0.000	0.000	0.000	0.000	0.000	0.000
		0.476	14.350	0.000	0.000	0.000	0.000	0.000	0.000
		0.740	12.038	0.608	0.107	0.020	0.004	0.000	0.000
	λ_{10}	0.127	8.891	0.000	0.000	0.000	0.000	0.000	0.000
		0.307	12.043	0.000	0.000	0.000	0.000	0.000	0.000
		0.476	12.728	0.000	0.000	0.000	0.000	0.000	0.000
		0.740	10.505	0.996	0.972	0.947	0.886	0.839	0.722
	Louvain CDA modularity	0.127	0.137	0.000	0.000	0.000	0.000	0.000	0.000
		0.307	0.074	0.000	0.000	0.000	0.000	0.000	0.000
		0.476	0.049	0.000	0.000	0.000	0.000	0.000	0.000
		0.740	0.024	0.981	0.900	0.787	0.716	0.610	0.440
	density	0.130	0.130	1.000	1.000	0.991	0.977	0.959	0.885
		0.314	0.314	0.998	0.986	0.964	0.930	0.886	0.802
		0.488	0.488	0.996	0.980	0.946	0.902	0.858	0.744
		0.759	0.759	0.993	0.968	0.927	0.887	0.840	0.741
	max betweenness centrality	0.127	0.009	0.750	0.271	0.119	0.079	0.067	0.040
		0.307	0.005	1.000	0.813	0.605	0.466	0.390	0.249
		0.476	0.003	0.997	0.852	0.761	0.639	0.533	0.379
		0.740	0.001	0.994	0.930	0.882	0.821	0.766	0.651
	transitivity	0.127	0.132	0.999	0.990	0.971	0.923	0.886	0.771
		0.307	0.318	0.995	0.970	0.931	0.893	0.856	0.749
		0.476	0.494	0.992	0.976	0.944	0.907	0.857	0.745
		0.740	0.768	0.993	0.970	0.943	0.884	0.824	0.728
	triangle density	0.127	0.002	0.999	0.991	0.968	0.910	0.844	0.769
		0.307	0.032	0.995	0.975	0.948	0.917	0.871	0.783
		0.476	0.119	0.994	0.981	0.944	0.903	0.854	0.755
		0.740	0.446	0.989	0.971	0.934	0.893	0.836	0.731
500	λ_1	0.071	37.783	0.966	0.846	0.712	0.602	0.504	0.325
		0.230	119.330	0.998	0.977	0.933	0.907	0.859	0.787
		0.413	213.258	0.999	0.980	0.947	0.904	0.864	0.781
		0.740	381.727	0.999	0.960	0.926	0.879	0.833	0.752
	λ_3	0.071	11.362	0.000	0.000	0.000	0.000	0.000	0.000
		0.230	18.167	0.000	0.000	0.000	0.000	0.000	0.000
		0.413	20.806	0.000	0.000	0.000	0.000	0.000	0.000
		0.740	17.709	0.000	0.000	0.000	0.000	0.000	0.000
	λ_{10}	0.071	10.551	0.000	0.000	0.000	0.000	0.000	0.000
		0.230	16.902	0.000	0.000	0.000	0.000	0.000	0.000
		0.413	19.377	0.000	0.000	0.000	0.000	0.000	0.000
		0.740	16.336	0.000	0.000	0.000	0.000	0.000	0.000
	Louvain CDA modularity	0.071	0.135	0.000	0.000	0.000	0.000	0.000	0.000
		0.230	0.066	0.000	0.000	0.000	0.000	0.000	0.000
		0.413	0.041	0.000	0.000	0.000	0.000	0.000	0.000
		0.740	0.018	0.671	0.153	0.038	0.019	0.005	0.002
	density	0.073	0.073	1.000	0.998	0.990	0.975	0.946	0.882
		0.236	0.236	0.999	0.992	0.977	0.938	0.905	0.839
		0.423	0.423	0.999	0.984	0.953	0.915	0.885	0.791
		0.759	0.759	0.998	0.965	0.921	0.883	0.845	0.755
	max betweenness centrality	0.071	0.005	0.097	0.031	0.012	0.009	0.006	0.002
		0.230	0.003	0.959	0.643	0.462	0.352	0.269	0.190
		0.413	0.002	0.995	0.902	0.764	0.630	0.517	0.344
		0.740	0.001	0.988	0.943	0.891	0.846	0.791	0.671
	transitivity	0.071	0.074	1.000	0.999	0.979	0.953	0.887	0.772
		0.230	0.239	0.998	0.987	0.952	0.906	0.872	0.792
		0.413	0.428	0.999	0.986	0.959	0.921	0.882	0.808
		0.740	0.769	0.997	0.962	0.930	0.897	0.835	0.754
	triangle density	0.071	0.000	0.998	0.977	0.911	0.849	0.809	0.668
		0.230	0.013	0.998	0.988	0.957	0.927	0.890	0.811
		0.413	0.077	0.999	0.983	0.956	0.920	0.862	0.797
		0.740	0.446	0.999	0.957	0.928	0.885	0.833	0.754

Table 2.2: Confidence interval coverage for different densities based on Monte Carlo simulations using the high density generating function when $S = 1000$.

			average for	Proportion of bootstrap CI that cover truth for					
n	statistic	ρ_n	true graphs	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
750	λ_1	0.008	7.696	0.981	0.899	0.700	0.563	0.446	0.311
		0.029	25.866	0.991	0.936	0.890	0.841	0.786	0.673
		0.058	49.481	0.985	0.944	0.880	0.828	0.794	0.696
		0.113	95.648	0.983	0.946	0.897	0.829	0.779	0.681
	λ_3	0.008	5.338	0.000	0.000	0.000	0.000	0.000	0.000
		0.029	12.664	0.758	0.567	0.431	0.333	0.279	0.209
		0.058	22.937	0.979	0.898	0.825	0.764	0.709	0.619
		0.113	43.090	0.984	0.902	0.833	0.759	0.705	0.610
	λ_{10}	0.008	4.773	0.698	0.171	0.052	0.017	0.012	0.004
		0.029	8.859	0.000	0.000	0.000	0.000	0.000	0.000
		0.058	11.656	0.000	0.000	0.000	0.000	0.000	0.000
		0.113	14.127	0.989	0.937	0.876	0.831	0.793	0.701
	Louvain CDA modularity	0.008	0.436	0.000	0.000	0.000	0.000	0.000	0.000
		0.029	0.323	0.891	0.806	0.712	0.646	0.585	0.494
		0.058	0.307	0.974	0.915	0.836	0.756	0.689	0.589
		0.113	0.294	0.988	0.948	0.903	0.842	0.754	0.673
	density	0.008	0.008	0.007	0.000	0.000	0.000	0.000	0.000
		0.029	0.029	0.952	0.719	0.569	0.413	0.317	0.203
		0.058	0.058	0.980	0.932	0.876	0.821	0.758	0.678
		0.113	0.113	0.993	0.948	0.905	0.855	0.802	0.710
	max betweenness centrality	0.008	0.023	0.806	0.334	0.176	0.125	0.094	0.062
		0.029	0.008	1.000	0.990	0.885	0.797	0.692	0.508
		0.058	0.006	1.000	0.969	0.869	0.713	0.586	0.464
		0.113	0.005	1.000	0.996	0.960	0.890	0.811	0.665
	transitivity	0.008	0.010	1.000	1.000	1.000	0.998	0.995	0.973
		0.029	0.039	0.987	0.946	0.901	0.848	0.794	0.714
		0.058	0.076	0.983	0.945	0.884	0.840	0.809	0.699
		0.113	0.148	0.976	0.926	0.870	0.824	0.782	0.668
	triangle density	0.008	0.000	1.000	0.998	0.992	0.972	0.935	0.869
		0.029	0.000	0.998	0.961	0.920	0.883	0.835	0.743
		0.058	0.000	0.992	0.930	0.894	0.846	0.800	0.709
		0.113	0.002	0.976	0.930	0.876	0.824	0.771	0.682
1000	λ_1	0.006	8.002	0.977	0.788	0.611	0.472	0.332	0.204
		0.026	30.381	0.998	0.941	0.894	0.844	0.787	0.689
		0.054	61.907	0.998	0.942	0.886	0.831	0.792	0.694
		0.113	127.414	0.996	0.950	0.900	0.844	0.781	0.672
	λ_3	0.006	5.528	0.000	0.000	0.000	0.000	0.000	0.000
		0.026	14.699	0.903	0.682	0.563	0.461	0.416	0.333
		0.054	28.486	0.986	0.947	0.891	0.846	0.789	0.678
		0.113	57.194	0.989	0.934	0.887	0.836	0.776	0.694
	λ_{10}	0.006	5.000	0.141	0.001	0.000	0.000	0.000	0.000
		0.026	9.885	0.000	0.000	0.000	0.000	0.000	0.000
		0.054	13.380	0.000	0.000	0.000	0.000	0.000	0.000
		0.113	16.554	0.995	0.972	0.914	0.848	0.801	0.697
	Louvain CDA modularity	0.006	0.425	0.000	0.000	0.000	0.000	0.000	0.000
		0.026	0.319	0.958	0.898	0.834	0.786	0.731	0.604
		0.054	0.302	0.971	0.881	0.804	0.747	0.643	0.557
		0.113	0.291	0.998	0.969	0.924	0.875	0.834	0.707
	density	0.006	0.006	0.000	0.000	0.000	0.000	0.000	0.000
		0.026	0.026	0.810	0.537	0.366	0.285	0.173	0.101
		0.054	0.054	0.980	0.895	0.828	0.776	0.735	0.605
		0.113	0.113	0.988	0.966	0.909	0.869	0.793	0.690
	max betweenness centrality	0.006	0.018	0.658	0.210	0.104	0.066	0.045	0.028
		0.026	0.006	1.000	0.995	0.897	0.776	0.697	0.507
		0.054	0.004	1.000	0.936	0.765	0.643	0.522	0.370
		0.113	0.003	1.000	0.997	0.918	0.874	0.843	0.706
	transitivity	0.006	0.008	1.000	1.000	1.000	0.998	0.993	0.968
		0.026	0.034	0.993	0.957	0.898	0.846	0.794	0.685
		0.054	0.071	0.991	0.945	0.888	0.815	0.763	0.654
		0.113	0.147	0.992	0.944	0.868	0.800	0.747	0.625
	triangle density	0.006	0.000	1.000	0.996	0.974	0.946	0.906	0.800
		0.026	0.000	0.998	0.980	0.937	0.872	0.818	0.716
		0.054	0.000	0.993	0.957	0.875	0.832	0.765	0.649
		0.113	0.002	0.994	0.943	0.886	0.805	0.737	0.606

Table 2.3: Confidence interval coverage for different densities based on Monte Carlo simulations using the horseshoe generating function when $S = 1000$.

			average for		Proportion of bootstrap CI that cover truth for					
n	average $\hat{a}$	statistic	true graphs	S	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
300	0.000105	density	0.187	100	1.000	0.990	0.930	0.880	0.880	0.760
				200	0.995	0.950	0.910	0.870	0.840	0.770
				300	0.993	0.947	0.910	0.850	0.803	0.740
				500	0.996	0.950	0.916	0.854	0.806	0.726
				750	0.996	0.943	0.897	0.835	0.795	0.709
				1000	0.996	0.939	0.897	0.847	0.812	0.712
				1250	0.993	0.941	0.902	0.845	0.811	0.714
				1500	0.995	0.945	0.901	0.846	0.807	0.713
				2000	0.995	0.949	0.900	0.849	0.809	0.719
		λ_1	75.206	100	1.000	0.990	0.920	0.890	0.840	0.700
				200	0.995	0.955	0.905	0.870	0.815	0.730
				300	0.993	0.950	0.890	0.853	0.820	0.733
				500	0.996	0.944	0.894	0.860	0.824	0.722
				750	0.996	0.941	0.892	0.843	0.819	0.719
				1000	0.994	0.942	0.898	0.843	0.813	0.725
				1250	0.993	0.941	0.897	0.850	0.818	0.721
				1500	0.993	0.949	0.899	0.853	0.815	0.724
				2000	0.995	0.952	0.899	0.851	0.815	0.719
		transitivity	0.332	100	1.000	0.980	0.940	0.890	0.810	0.700
				200	1.000	0.965	0.925	0.890	0.825	0.680
				300	0.997	0.947	0.917	0.870	0.817	0.700
				500	0.992	0.946	0.896	0.836	0.796	0.702
				750	0.991	0.949	0.904	0.841	0.815	0.715
				1000	0.990	0.946	0.888	0.847	0.812	0.712
				1250	0.990	0.946	0.890	0.839	0.806	0.712
				1500	0.989	0.946	0.893	0.845	0.809	0.718
				2000	0.990	0.949	0.905	0.848	0.805	0.712
500	0.000059	density	0.188	100	0.950	0.910	0.890	0.830	0.790	0.640
				200	0.975	0.925	0.890	0.845	0.790	0.630
				300	0.983	0.953	0.897	0.850	0.803	0.663
				500	0.982	0.954	0.914	0.848	0.804	0.692
				750	0.983	0.948	0.897	0.849	0.809	0.707
				1000	0.982	0.949	0.905	0.852	0.812	0.712
				1250	0.989	0.957	0.917	0.865	0.825	0.713
				1500	0.991	0.958	0.919	0.876	0.829	0.715
				2000	0.993	0.960	0.919	0.869	0.827	0.722
		λ_1	125.416	100	0.940	0.910	0.860	0.790	0.780	0.660
				200	0.990	0.920	0.890	0.840	0.790	0.670
				300	0.990	0.947	0.910	0.867	0.817	0.670
				500	0.986	0.954	0.928	0.858	0.828	0.668
				750	0.991	0.951	0.913	0.863	0.815	0.696
				1000	0.988	0.951	0.915	0.866	0.811	0.709
				1250	0.990	0.954	0.914	0.872	0.820	0.716
				1500	0.991	0.955	0.919	0.875	0.817	0.719
				2000	0.995	0.959	0.920	0.881	0.816	0.716
		transitivity	0.333	100	0.990	0.960	0.800	0.770	0.710	0.630
				200	0.990	0.950	0.935	0.820	0.775	0.660
				300	0.987	0.953	0.930	0.863	0.803	0.717
				500	0.986	0.964	0.926	0.874	0.814	0.706
				750	0.988	0.969	0.931	0.875	0.825	0.727
				1000	0.982	0.957	0.922	0.865	0.812	0.722
				1250	0.993	0.962	0.921	0.862	0.815	0.720
				1500	0.990	0.961	0.917	0.862	0.815	0.713
				2000	0.997	0.963	0.923	0.874	0.824	0.713

Table 2.4: Confidence interval coverage for different number of simulated true graphs S based on Monte Carlo simulations using the product generating function when $n = 300$ or $n = 500$ and $\rho_n = 0.1874$.

			average for		Proportion of bootstrap CI that cover truth for					
n	average $\hat{a}$	statistic	true graphs	S	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
300	0.000046	density	0.112	100	0.990	0.990	0.960	0.960	0.950	0.770
				200	0.990	0.980	0.955	0.945	0.925	0.850
				300	0.990	0.983	0.963	0.940	0.903	0.827
				500	0.992	0.976	0.952	0.922	0.880	0.802
				750	0.992	0.972	0.948	0.905	0.860	0.780
				1000	0.994	0.974	0.945	0.901	0.866	0.781
				1250	0.995	0.975	0.943	0.898	0.858	0.775
				1500	0.995	0.975	0.937	0.899	0.859	0.774
				2000	0.998	0.973	0.940	0.901	0.855	0.761
		λ_2	25.306	100	0.970	0.960	0.810	0.720	0.700	0.630
				200	0.980	0.965	0.910	0.875	0.800	0.665
				300	0.990	0.963	0.930	0.900	0.830	0.680
				500	0.994	0.954	0.910	0.876	0.818	0.662
				750	0.991	0.953	0.905	0.861	0.800	0.669
				1000	0.995	0.959	0.911	0.859	0.805	0.669
				1250	0.995	0.957	0.912	0.866	0.819	0.674
				1500	0.995	0.965	0.917	0.871	0.821	0.682
				2000	0.997	0.963	0.919	0.869	0.819	0.693
		transitivity	0.146	100	0.980	0.850	0.810	0.780	0.750	0.670
				200	1.000	0.910	0.845	0.790	0.755	0.685
				300	0.987	0.873	0.820	0.793	0.743	0.663
				500	0.986	0.912	0.846	0.802	0.772	0.690
				750	0.983	0.920	0.856	0.809	0.772	0.667
				1000	0.995	0.932	0.881	0.813	0.777	0.674
				1250	0.994	0.934	0.885	0.818	0.779	0.674
				1500	0.995	0.943	0.889	0.830	0.791	0.685
				2000	0.994	0.945	0.885	0.822	0.776	0.681
500	0.000031	density	0.112	100	0.970	0.910	0.860	0.830	0.810	0.750
				200	0.985	0.955	0.890	0.855	0.815	0.750
				300	0.990	0.953	0.937	0.873	0.820	0.773
				500	1.000	0.962	0.946	0.892	0.836	0.746
				750	0.999	0.961	0.949	0.896	0.848	0.761
				1000	0.998	0.961	0.936	0.878	0.832	0.748
				1250	0.997	0.966	0.940	0.890	0.843	0.746
				1500	0.996	0.966	0.937	0.886	0.831	0.731
				2000	0.998	0.964	0.931	0.883	0.834	0.743
		λ_2	41.732	100	0.990	0.940	0.900	0.830	0.770	0.710
				200	1.000	0.970	0.905	0.795	0.765	0.695
				300	0.997	0.943	0.887	0.817	0.783	0.717
				500	0.998	0.948	0.910	0.844	0.802	0.734
				750	0.999	0.955	0.897	0.829	0.797	0.727
				1000	0.998	0.960	0.905	0.834	0.801	0.725
				1250	0.998	0.962	0.900	0.842	0.806	0.731
				1500	0.998	0.958	0.896	0.834	0.797	0.715
				2000	0.998	0.956	0.905	0.848	0.804	0.726
		transitivity	0.147	100	0.990	0.950	0.890	0.860	0.750	0.690
				200	0.985	0.960	0.940	0.880	0.845	0.720
				300	0.983	0.947	0.930	0.850	0.807	0.700
				500	0.988	0.958	0.908	0.838	0.790	0.670
				750	0.989	0.953	0.896	0.843	0.795	0.676
				1000	0.989	0.955	0.898	0.855	0.812	0.697
				1250	0.990	0.958	0.904	0.859	0.813	0.701
				1500	0.991	0.951	0.901	0.859	0.806	0.703
				2000	0.991	0.956	0.890	0.846	0.792	0.684

Table 2.5: Confidence interval coverage for different number of simulated true graphs S based on Monte Carlo simulations using the horseshoe generating function when $\rho_n = 0.1125$.

		average for		Proportion of bootstrap CI that cover truth for					
ρ_n	statistic	true graphs	S	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
0.569	density	0.570	100	1.000	0.990	0.970	0.950	0.910	0.810
			200	1.000	0.980	0.955	0.930	0.855	0.760
			300	1.000	0.980	0.947	0.913	0.843	0.757
			500	1.000	0.980	0.946	0.890	0.832	0.754
			750	1.000	0.975	0.925	0.855	0.815	0.715
			1000	0.996	0.975	0.928	0.858	0.821	0.733
			1250	0.998	0.970	0.918	0.851	0.815	0.718
			1500	0.997	0.969	0.924	0.863	0.820	0.723
			2000	0.994	0.970	0.924	0.862	0.820	0.722
	λ_1	286.543	100	1.000	0.940	0.800	0.740	0.730	0.590
			200	0.985	0.960	0.905	0.840	0.770	0.650
			300	0.983	0.940	0.910	0.840	0.750	0.623
			500	0.984	0.940	0.898	0.822	0.742	0.638
			750	0.992	0.948	0.907	0.841	0.792	0.687
			1000	0.993	0.944	0.903	0.842	0.784	0.701
			1250	0.996	0.962	0.908	0.848	0.796	0.700
			1500	0.997	0.963	0.907	0.862	0.797	0.699
			2000	0.998	0.966	0.910	0.868	0.814	0.708
	transitivity	0.576	100	1.000	0.920	0.870	0.730	0.650	0.580
			200	0.985	0.955	0.900	0.840	0.795	0.670
			300	0.983	0.927	0.903	0.803	0.753	0.630
			500	0.986	0.940	0.886	0.834	0.760	0.652
			750	0.999	0.949	0.908	0.856	0.795	0.688
			1000	0.998	0.948	0.901	0.855	0.800	0.699
			1250	0.998	0.967	0.909	0.866	0.806	0.698
			1500	0.997	0.967	0.911	0.869	0.807	0.691
			2000	0.998	0.969	0.912	0.869	0.821	0.703
0.759	density	0.759	100	0.990	0.990	0.920	0.920	0.910	0.770
			200	0.995	0.970	0.925	0.900	0.840	0.690
			300	0.990	0.977	0.917	0.870	0.840	0.723
			500	0.988	0.974	0.918	0.882	0.842	0.712
			750	0.996	0.977	0.928	0.893	0.849	0.747
			1000	0.990	0.970	0.926	0.895	0.861	0.748
			1250	0.992	0.965	0.928	0.890	0.855	0.739
			1500	0.990	0.958	0.915	0.873	0.827	0.707
			2000	0.990	0.957	0.906	0.861	0.818	0.700
	λ_1	381.690	100	0.960	0.940	0.860	0.850	0.800	0.680
			200	0.970	0.945	0.900	0.895	0.870	0.675
			300	0.987	0.957	0.910	0.897	0.813	0.710
			500	0.992	0.964	0.918	0.882	0.834	0.732
			750	0.987	0.967	0.907	0.863	0.832	0.737
			1000	0.989	0.962	0.912	0.874	0.833	0.739
			1250	0.990	0.958	0.903	0.845	0.809	0.710
			1500	0.990	0.961	0.908	0.850	0.813	0.711
			2000	0.989	0.957	0.902	0.849	0.814	0.708
	transitivity	0.768	100	0.950	0.940	0.870	0.840	0.760	0.690
			200	0.965	0.930	0.900	0.885	0.830	0.700
			300	0.983	0.957	0.910	0.890	0.850	0.693
			500	0.992	0.968	0.914	0.886	0.854	0.722
			750	0.987	0.967	0.915	0.881	0.837	0.735
			1000	0.988	0.962	0.917	0.878	0.836	0.734
			1250	0.990	0.956	0.896	0.854	0.802	0.706
			1500	0.989	0.958	0.903	0.860	0.805	0.711
			2000	0.990	0.957	0.900	0.855	0.807	0.711

Table 2.6: Confidence interval coverage for different number of simulated true graphs S based on Monte Carlo simulations using the high density generating function when $n = 500$.

		average for		Proportion of bootstrap CI that cover truth for					
n	statistic	true graphs	method	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
100	λ_1	17.222	HK1	0.998	0.982	0.940	0.895	0.854	0.775
			HK2	0.998	0.986	0.939	0.894	0.861	0.769
			HNN1	0.947	0.777	0.648	0.590	0.507	0.397
			emp	0.999	0.953	0.912	0.877	0.842	0.742
	λ_2	6.418	HK1	0.974	0.858	0.704	0.599	0.504	0.341
			HK2	1.000	0.947	0.879	0.796	0.679	0.514
			HNN1	0.966	0.896	0.771	0.675	0.612	0.537
			emp	0.343	0.057	0.012	0.004	0.002	0.000
	transitivity	0.220	HK1	0.999	0.992	0.971	0.938	0.907	0.847
			HK2	0.998	0.984	0.963	0.933	0.901	0.831
			HNN1	0.986	0.937	0.838	0.785	0.739	0.665
			emp	0.992	0.966	0.897	0.839	0.800	0.703
300	λ_1	50.535	HK1	0.995	0.959	0.918	0.876	0.836	0.739
			HK2	0.994	0.958	0.923	0.889	0.859	0.765
			HNN1	0.936	0.847	0.724	0.647	0.597	0.509
			emp	0.990	0.953	0.900	0.859	0.824	0.733
	λ_2	11.778	HK1	0.000	0.000	0.000	0.000	0.000	0.000
			HK2	0.001	0.000	0.000	0.000	0.000	0.000
			HNN1	0.978	0.915	0.857	0.814	0.754	0.651
			emp	0.000	0.000	0.000	0.000	0.000	0.000
	transitivity	0.221	HK1	0.997	0.982	0.940	0.897	0.859	0.759
			HK2	0.996	0.984	0.954	0.904	0.860	0.782
			HNN1	0.981	0.919	0.872	0.823	0.747	0.629
			emp	0.988	0.947	0.897	0.847	0.781	0.678
500	λ_1	83.925	HK1	0.983	0.955	0.904	0.849	0.782	0.680
			HK2	0.986	0.944	0.902	0.858	0.791	0.701
			HNN1	0.950	0.860	0.773	0.682	0.633	0.534
			emp	0.981	0.937	0.888	0.851	0.817	0.708
	λ_2	15.448	HK1	0.000	0.000	0.000	0.000	0.000	0.000
			HK2	0.000	0.000	0.000	0.000	0.000	0.000
			HNN1	0.982	0.915	0.854	0.814	0.753	0.658
			emp	0.000	0.000	0.000	0.000	0.000	0.000
	transitivity	0.222	HK1	0.986	0.945	0.899	0.867	0.814	0.708
			HK2	0.987	0.956	0.911	0.851	0.812	0.734
			HNN1	0.982	0.920	0.838	0.780	0.734	0.630
			emp	0.982	0.968	0.905	0.837	0.772	0.692

Table 2.7: Confidence interval coverage based on Monte Carlo simulations for different bootstrap methods for the true graphs from the product generating function with density $\rho_n = 0.125$, $S = 1000$ true graphs, with sample size n ranging from 100 to 500. The methods are: HK1 (our main method based on $\hat{h} \equiv \hat{h}^{(K1)}$ with $\hat{a}$), HK2 (our bootstrap method but using the linking function estimator $\hat{h}^{(K2)}$ with $a^{(optK2)}$ based on $\hat{h}^{(K2)}$), HNN1 (our bootstrap method but but using the linking function estimator $\hat{h}^{(NN1)}$ from Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size), emp (empirical bootstrap from Green and Shalizi (2022)), dot_prod_k (the bootstrap method from Levin and Levina (2019) based on assuming a k-dimensional ξ_i).

				Proportion of bootstrap CI that cover truth for					
				α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
statistic	n	average for true graphs	method						
density	250	0.160767	HK1	0.991	0.959	0.914	0.874	0.823	0.736
			HK2	0.993	0.956	0.908	0.868	0.826	0.721
			HNN1	0.981	0.914	0.873	0.818	0.762	0.636
			LLS_L	1.000	1.000	1.000	0.999	0.999	0.990
			asymp estimated var	0.993	0.973	0.948	0.909	0.864	0.781
			asymp infeasible var	0.990	0.943	0.895	0.846	0.795	0.701
			dot_prod_1	1.000	0.999	0.998	0.990	0.973	0.913
			dot_prod_3	1.000	0.999	0.998	0.991	0.967	0.906
			emp	0.990	0.961	0.910	0.871	0.828	0.742
	500	0.139211	HK1	0.993	0.955	0.889	0.827	0.771	0.672
			HK2	0.993	0.954	0.872	0.816	0.753	0.667
			HNN1	0.987	0.896	0.795	0.730	0.675	0.590
			asymp estimated var	0.998	0.978	0.948	0.907	0.868	0.770
			asymp infeasible var	0.993	0.952	0.900	0.855	0.797	0.676
			dot_prod_1	1.000	1.000	0.996	0.988	0.979	0.920
			dot_prod_3	1.000	1.000	0.997	0.991	0.985	0.936
			emp	0.991	0.942	0.891	0.841	0.784	0.677
	750	0.127979	HK1	0.979	0.937	0.877	0.842	0.788	0.684
			HK2	0.973	0.928	0.873	0.824	0.774	0.665
			HNN1	0.944	0.866	0.775	0.721	0.659	0.556
			asymp estimated var	0.994	0.980	0.953	0.914	0.883	0.799
			asymp infeasible var	0.989	0.948	0.901	0.852	0.803	0.704
			dot_prod_1	1.000	0.999	0.993	0.982	0.964	0.920
			dot_prod_3	1.000	0.999	0.993	0.982	0.965	0.923
			emp	0.978	0.933	0.887	0.835	0.795	0.691
triangle density	250	0.009854	HK1	0.990	0.969	0.942	0.886	0.852	0.745
			HK2	0.990	0.974	0.941	0.902	0.869	0.749
			HNN1	0.990	0.946	0.902	0.839	0.788	0.675
			LLS_L	1.000	0.999	0.995	0.983	0.968	0.925
			LLS_Q	1.000	0.999	0.995	0.983	0.966	0.923
			asymp infeasible var	0.987	0.953	0.905	0.852	0.794	0.697
			dot_prod_1	1.000	0.999	0.997	0.983	0.975	0.930
			dot_prod_3	1.000	0.999	0.999	0.983	0.976	0.933
			emp	0.994	0.966	0.935	0.871	0.821	0.732
	500	0.006402	HK1	0.991	0.959	0.902	0.840	0.794	0.692
			HK2	0.997	0.960	0.896	0.847	0.790	0.700
			HNN1	0.986	0.929	0.826	0.760	0.704	0.606
			asymp infeasible var	0.993	0.951	0.889	0.852	0.797	0.695
			dot_prod_1	1.000	1.000	0.997	0.987	0.970	0.911
			dot_prod_3	1.000	1.000	0.998	0.993	0.973	0.921
			emp	0.996	0.951	0.882	0.839	0.789	0.692
	750	0.004970	HK1	0.987	0.946	0.898	0.854	0.805	0.695
			HK2	0.987	0.943	0.892	0.852	0.793	0.697
			HNN1	0.962	0.904	0.820	0.752	0.703	0.598
			asymp infeasible var	0.987	0.940	0.900	0.852	0.803	0.703
			dot_prod_1	1.000	1.000	0.995	0.985	0.965	0.919
			dot_prod_3	1.000	1.000	0.995	0.985	0.967	0.922
			emp	0.977	0.938	0.900	0.841	0.804	0.709

Table 2.8: Confidence interval coverage based on Monte Carlo simulations for different bootstrap methods for the true graphs from the product generating function with density $\rho_n \sim \sqrt[4]{\frac{\log(n)}{n}}$, $S = 1000$ true graphs. The methods are: HK1 (our main method based on $\hat{h} \equiv \hat{h}^{(K1)}$ with $\hat{a}$), HK2 (our bootstrap method but using the linking function estimator $\hat{h}^{(K2)}$ with $a^{(optK2)}$ based on $\hat{h}^{(K2)}$), HNN1 (our bootstrap method but but using the linking function estimator $\hat{h}^{(NN1)}$ from Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size), emp (empirical bootstrap from Green and Shalizi (2022)), dot_prod_k (the bootstrap method from Levin and Levina (2019) based on assuming a k-dimensional ξ_i), asymp estimated var (the asymptotic distribution from Bickel, Chen, and Levina (2011) with variance estimated according to the formula in Green and Shalizi (2022)), asymp infeasible var (the asymptotic distribution from Bickel, Chen, and Levina (2011) with the true theoretical variance), LLS_L and LLS_Q (the linear and quadratic methods from Lin, Lunde, and Sarkar (2020)).

		average for		Proportion of bootstrap CI that cover truth for					
ρ_n	statistic	true graphs	method	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
0.056	density	0.056	HK1	0.989	0.927	0.864	0.786	0.721	0.634
			HK2	0.968	0.825	0.654	0.514	0.419	0.346
			HNN1	0.253	0.066	0.034	0.017	0.005	0.001
			dot_prod_1	0.005	0.000	0.000	0.000	0.000	0.000
			dot_prod_3	1.000	0.998	0.985	0.973	0.955	0.901
			emp	0.999	0.985	0.960	0.920	0.875	0.800
	λ_1	32.530	HK1	0.988	0.968	0.931	0.866	0.820	0.722
			HK2	0.996	0.977	0.945	0.898	0.849	0.767
			HNN1	0.925	0.751	0.631	0.522	0.440	0.332
			emp	0.979	0.925	0.871	0.813	0.763	0.669
	λ_2	21.481	HK1	0.997	0.957	0.903	0.860	0.827	0.729
			HK2	0.991	0.911	0.807	0.748	0.686	0.589
			HNN1	0.970	0.845	0.696	0.585	0.507	0.409
			emp	0.935	0.811	0.694	0.631	0.560	0.480
	transitivity	0.073	HK1	0.997	0.938	0.890	0.848	0.800	0.672
			HK2	0.970	0.930	0.867	0.826	0.781	0.688
			HNN1	0.966	0.901	0.818	0.767	0.702	0.590
			emp	0.993	0.956	0.904	0.851	0.786	0.692
0.084	density	0.084	HK1	0.990	0.953	0.914	0.874	0.833	0.722
			HK2	0.990	0.950	0.909	0.861	0.794	0.650
			HNN1	0.910	0.854	0.732	0.619	0.563	0.436
			dot_prod_1	0.000	0.000	0.000	0.000	0.000	0.000
			dot_prod_3	0.999	0.996	0.987	0.970	0.948	0.903
			emp	0.996	0.986	0.957	0.914	0.875	0.797
	λ_1	48.240	HK1	0.997	0.971	0.912	0.863	0.817	0.705
			HK2	0.998	0.976	0.934	0.875	0.821	0.721
			HNN1	0.973	0.919	0.843	0.809	0.760	0.638
			emp	0.997	0.967	0.900	0.837	0.786	0.681
	λ_2	31.542	HK1	0.996	0.971	0.931	0.876	0.827	0.730
			HK2	0.993	0.951	0.896	0.854	0.797	0.672
			HNN1	0.989	0.959	0.902	0.834	0.805	0.694
			emp	0.991	0.941	0.876	0.824	0.780	0.663
	transitivity	0.110	HK1	0.997	0.949	0.886	0.825	0.779	0.669
			HK2	0.996	0.950	0.896	0.845	0.796	0.702
			HNN1	0.955	0.885	0.814	0.761	0.688	0.578
			emp	0.998	0.965	0.922	0.874	0.819	0.738
0.113	density	0.113	HK1	0.993	0.975	0.922	0.866	0.823	0.732
			HK2	0.989	0.970	0.914	0.869	0.829	0.718
			HNN1	0.990	0.937	0.875	0.827	0.777	0.661
			dot_prod_1	0.000	0.000	0.000	0.000	0.000	0.000
			dot_prod_3	1.000	0.993	0.985	0.975	0.958	0.898
			emp	0.993	0.982	0.955	0.918	0.870	0.792
	λ_1	63.879	HK1	0.997	0.968	0.908	0.856	0.781	0.662
			HK2	1.000	0.973	0.916	0.874	0.807	0.690
			HNN1	0.989	0.937	0.879	0.789	0.738	0.654
			emp	0.993	0.958	0.899	0.839	0.798	0.693
	λ_2	41.713	HK1	0.994	0.961	0.909	0.855	0.810	0.694
			HK2	0.994	0.941	0.898	0.845	0.793	0.697
			HNN1	0.988	0.955	0.901	0.857	0.799	0.686
			emp	0.994	0.945	0.883	0.814	0.749	0.643
	transitivity	0.147	HK1	0.996	0.956	0.895	0.820	0.766	0.672
			HK2	1.000	0.957	0.902	0.837	0.784	0.661
			HNN1	0.956	0.846	0.771	0.694	0.655	0.550
			emp	0.989	0.944	0.891	0.846	0.792	0.682

Table 2.9: Confidence interval coverage based on Monte Carlo simulations for different bootstrap methods for the true graphs from the horseshoe generating function with sample size $n = 500$, $S = 1000$ true graphs, and different values of density ρ_n ranging from 0.056 to 0.1125. The methods are: HK1 (our main method based on $\hat{h} \equiv \hat{h}^{(K1)}$ with $\hat{a}$), HK2 (our bootstrap method but using the linking function estimator $\hat{h}^{(K2)}$ with $a^{(optK2)}$ based on $\hat{h}^{(K2)}$), HNN1 (our bootstrap method but but using the linking function estimator $\hat{h}^{(NN1)}$ from Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size), emp (empirical bootstrap from Green and Shalizi (2022)), dot_prod_k (the bootstrap method from Levin and Levina (2019) based on assuming a k-dimensional ξ_i).

		average for		Proportion of bootstrap CI that cover truth for					
ρ_n	statistic	true graphs	method	α =0.01	α =0.05	α =0.1	α =0.15	α =0.2	α =0.3
0.380	density	0.380	HK1	0.998	0.984	0.964	0.928	0.879	0.800
			HK2	0.995	0.978	0.954	0.931	0.888	0.795
			HNN1	0.982	0.921	0.859	0.783	0.735	0.643
			dot_prod_1	1.000	0.996	0.987	0.975	0.948	0.873
			dot_prod_3	1.000	0.996	0.986	0.973	0.944	0.873
			emp	0.998	0.983	0.948	0.932	0.885	0.772
	λ_1	114.861	HK1	0.994	0.964	0.937	0.892	0.851	0.777
			HK2	0.999	0.970	0.939	0.907	0.869	0.793
			HNN1	0.990	0.973	0.926	0.878	0.834	0.728
			emp	1.000	0.984	0.957	0.935	0.902	0.817
	λ_2	15.962	HK1	0.000	0.000	0.000	0.000	0.000	0.000
			HK2	0.000	0.000	0.000	0.000	0.000	0.000
			HNN1	1.000	0.992	0.980	0.952	0.929	0.847
			emp	0.000	0.000	0.000	0.000	0.000	0.000
	transitivity	0.384	HK1	0.994	0.973	0.937	0.902	0.868	0.786
			HK2	1.000	0.981	0.937	0.897	0.859	0.775
			HNN1	0.997	0.973	0.948	0.915	0.880	0.774
			emp	0.998	0.973	0.888	0.839	0.778	0.694
0.570	density	0.570	HK1	0.994	0.969	0.945	0.904	0.849	0.756
			HK2	0.993	0.973	0.923	0.882	0.827	0.713
			HNN1	0.994	0.964	0.926	0.889	0.836	0.731
			dot_prod_1	1.000	0.996	0.991	0.973	0.947	0.886
			dot_prod_3	1.000	0.996	0.987	0.963	0.939	0.857
			emp	0.991	0.945	0.915	0.863	0.814	0.723
	λ_1	171.787	HK1	0.999	0.978	0.941	0.905	0.852	0.767
			HK2	0.999	0.978	0.941	0.900	0.857	0.779
			HNN1	0.990	0.972	0.939	0.884	0.838	0.740
			emp	0.990	0.971	0.937	0.887	0.850	0.757
	λ_2	15.963	HK1	0.000	0.000	0.000	0.000	0.000	0.000
			HK2	0.000	0.000	0.000	0.000	0.000	0.000
			HNN1	1.000	0.993	0.977	0.949	0.895	0.832
			emp	0.000	0.000	0.000	0.000	0.000	0.000
	transitivity	0.576	HK1	0.998	0.979	0.934	0.908	0.864	0.770
			HK2	0.997	0.978	0.940	0.902	0.867	0.776
			HNN1	0.987	0.969	0.920	0.877	0.830	0.730
			emp	0.985	0.938	0.891	0.819	0.787	0.652
0.759	density	0.759	HK1	0.994	0.963	0.911	0.872	0.841	0.737
			HK2	0.992	0.961	0.919	0.870	0.825	0.747
			HNN1	0.986	0.942	0.902	0.859	0.809	0.694
			dot_prod_1	1.000	0.998	0.990	0.975	0.945	0.881
			dot_prod_3	1.000	0.997	0.986	0.959	0.922	0.850
			emp	0.985	0.947	0.890	0.835	0.810	0.706
	λ_1	228.714	HK1	0.996	0.965	0.926	0.884	0.842	0.731
			HK2	0.993	0.959	0.926	0.877	0.841	0.734
			HNN1	0.991	0.962	0.931	0.865	0.819	0.702
			emp	0.983	0.953	0.914	0.864	0.813	0.702
	λ_2	13.730	HK1	0.020	0.000	0.000	0.000	0.000	0.000
			HK2	0.978	0.817	0.615	0.498	0.379	0.217
			HNN1	0.999	0.988	0.965	0.932	0.892	0.784
			emp	0.000	0.000	0.000	0.000	0.000	0.000
	transitivity	0.768	HK1	0.994	0.961	0.916	0.885	0.833	0.720
			HK2	0.993	0.961	0.916	0.886	0.829	0.732
			HNN1	0.992	0.965	0.932	0.864	0.815	0.712
			emp	0.970	0.920	0.872	0.814	0.776	0.643

Table 2.10: Confidence interval coverage based on Monte Carlo simulations for different bootstrap methods for the true graphs from the high density generating function with sample size $n = 300$, $S = 1000$ true graphs, and different values of density ρ_n ranging from 0.38 to 0.79. The methods are: HK1 (our main method based on $\hat{h} \equiv \hat{h}^{(K1)}$ with $\hat{a}$), HK2 (our bootstrap method but using the linking function estimator $\hat{h}^{(K2)}$ with $a^{(optK2)}$ based on $\hat{h}^{(K2)}$), HNN1 (our bootstrap method but but using the linking function estimator $\hat{h}^{(NN1)}$ from Zhang, Levina, and Zhu (2017) with their optimal choice of neighbourhood size), emp (empirical bootstrap from Green and Shalizi (2022)), dot_prod_k (the bootstrap method from Levin and Levina (2019) based on assuming a k-dimensional ξ_i).

			Proportion of bootstrap CI that cover truth for					
true value	c	average $\hat{a}$	α=0.01	α=0.05	α=0.1	α=0.15	α=0.2	α=0.3
0.1109	0.10	1.045577e-06	0.990	0.962	0.936	0.857	0.818	0.736
	0.25	2.613943e-06	0.997	0.987	0.949	0.913	0.862	0.783
	0.50	5.227885e-06	0.999	0.984	0.956	0.927	0.887	0.790
	0.75	7.841828e-06	0.995	0.976	0.950	0.912	0.865	0.767
	0.90	9.410194e-06	0.999	0.987	0.961	0.926	0.880	0.778
	1.00	1.045577e-05	0.996	0.989	0.961	0.947	0.912	0.834
	1.10	1.150135e-05	0.999	0.985	0.971	0.948	0.912	0.829
	1.25	1.306971e-05	0.998	0.979	0.947	0.913	0.842	0.743
	2	2.091154e-05	0.477	0.223	0.117	0.071	0.051	0.032
	4	4.182308e-05	0.000	0.000	0.000	0.000	0.000	0.000
	10	1.045577e-04	0.000	0.000	0.000	0.000	0.000	0.000
0.2218	0.10	3.524835e-06	0.987	0.941	0.886	0.849	0.791	0.704
	0.25	8.812088e-06	0.997	0.956	0.908	0.862	0.826	0.715
	0.50	1.762418e-05	0.996	0.973	0.926	0.903	0.853	0.771
	0.75	2.643626e-05	0.993	0.953	0.913	0.878	0.806	0.713
	0.90	3.172352e-05	0.984	0.955	0.930	0.903	0.869	0.769
	1.00	3.524835e-05	0.995	0.969	0.947	0.905	0.856	0.745
	1.10	3.877319e-05	0.991	0.959	0.941	0.891	0.850	0.743
	1.25	4.406044e-05	0.995	0.968	0.940	0.900	0.860	0.763
	2	7.049670e-05	0.992	0.939	0.883	0.819	0.768	0.691
	4	1.409934e-04	0.785	0.473	0.390	0.317	0.258	0.171
	10	3.524835e-04	0.012	0.003	0.002	0.000	0.000	0.000
0.3327	0.10	5.842118e-06	0.990	0.935	0.883	0.837	0.784	0.676
	0.25	1.460529e-05	0.974	0.935	0.881	0.816	0.769	0.667
	0.50	2.921059e-05	0.991	0.968	0.909	0.865	0.802	0.674
	0.75	4.381588e-05	0.992	0.966	0.927	0.893	0.834	0.722
	0.90	5.257906e-05	0.990	0.956	0.907	0.873	0.830	0.700
	1.00	5.842118e-05	0.991	0.958	0.921	0.874	0.829	0.716
	1.10	6.426330e-05	0.993	0.965	0.929	0.881	0.835	0.720
	1.25	7.302647e-05	0.993	0.955	0.899	0.868	0.818	0.690
	2	1.168424e-04	0.990	0.955	0.923	0.876	0.824	0.710
	4	2.336847e-04	0.969	0.917	0.854	0.808	0.739	0.647
	10	5.842118e-04	0.732	0.509	0.384	0.312	0.277	0.195
0.4436	0.10	7.561251e-06	0.995	0.955	0.903	0.852	0.788	0.708
	0.25	1.890313e-05	0.994	0.954	0.903	0.852	0.805	0.719
	0.50	3.780625e-05	0.990	0.949	0.894	0.857	0.823	0.722
	0.75	5.670938e-05	0.994	0.969	0.921	0.853	0.811	0.712
	0.90	6.805126e-05	0.993	0.962	0.900	0.862	0.839	0.763
	1.00	7.561251e-05	0.991	0.954	0.887	0.848	0.803	0.691
	1.10	8.317376e-05	0.986	0.948	0.899	0.861	0.820	0.715
	1.25	9.451564e-05	0.990	0.952	0.906	0.868	0.830	0.749
	2	1.512250e-04	0.989	0.937	0.884	0.849	0.806	0.715
	4	3.024500e-04	0.989	0.955	0.911	0.851	0.796	0.707
	10	7.561251e-04	0.951	0.875	0.792	0.731	0.673	0.573

Table 2.11: Confidence interval coverage for transitivity at different bandwidths $c \times \hat{a}$ and at different densities ρ_n, based on Monte Carlo simulations using the product generating function when $n = 500$ and $S = 1000$.

			Proportion of bootstrap CI that cover truth for					
ρ_n	c	average $\hat{a}$	α=0.01	α=0.05	α=0.1	α=0.15	α=0.2	α=0.3
0.028125	0.01	1.402691e-08	1.000	0.991	0.979	0.962	0.935	0.854
	0.10	1.402691e-07	1.000	0.988	0.976	0.960	0.931	0.843
	0.25	3.506728e-07	1.000	0.992	0.978	0.965	0.944	0.870
	0.50	7.013456e-07	0.999	0.989	0.970	0.949	0.918	0.842
	0.75	1.052018e-06	0.999	0.945	0.884	0.831	0.778	0.666
	0.90	1.262422e-06	0.986	0.831	0.733	0.570	0.450	0.317
	1.00	1.402691e-06	0.853	0.538	0.370	0.256	0.194	0.102
	1.10	1.542960e-06	0.687	0.325	0.174	0.105	0.064	0.025
	1.25	1.753364e-06	0.355	0.092	0.026	0.008	0.005	0.001
	2	2.805383e-06	0.000	0.000	0.000	0.000	0.000	0.000
	4	5.610765e-06	0.020	0.005	0.001	0.000	0.000	0.000
	10	1.402691e-05	0.853	0.706	0.593	0.519	0.457	0.379
	20	2.805383e-05	0.979	0.931	0.848	0.796	0.746	0.634
	100	1.402691e-04	0.981	0.949	0.887	0.829	0.755	0.658
0.056250	0.01	6.201768e-08	0.997	0.985	0.964	0.921	0.887	0.807
	0.10	6.201768e-07	0.997	0.984	0.957	0.929	0.877	0.799
	0.25	1.550442e-06	0.998	0.979	0.956	0.926	0.883	0.789
	0.50	3.100884e-06	0.997	0.988	0.966	0.931	0.897	0.808
	0.75	4.651326e-06	0.997	0.983	0.958	0.916	0.880	0.800
	0.90	5.581592e-06	0.996	0.969	0.930	0.856	0.810	0.704
	1.00	6.201768e-06	0.985	0.930	0.864	0.806	0.751	0.647
	1.10	6.821945e-06	0.986	0.876	0.799	0.720	0.654	0.543
	1.25	7.752210e-06	0.939	0.814	0.720	0.610	0.502	0.313
	2	1.240354e-05	0.835	0.465	0.268	0.185	0.126	0.067
	4	2.480707e-05	0.641	0.363	0.202	0.139	0.111	0.067
	10	6.201768e-05	0.767	0.638	0.525	0.466	0.428	0.343
	20	1.240354e-04	0.930	0.825	0.771	0.702	0.657	0.577
	100	6.201768e-04	0.969	0.871	0.802	0.745	0.698	0.599
0.084375	0.01	1.469518e-07	0.997	0.982	0.947	0.917	0.867	0.774
	0.10	1.469518e-06	0.992	0.977	0.947	0.910	0.867	0.764
	0.25	3.673796e-06	0.992	0.977	0.950	0.922	0.886	0.792
	0.50	7.347591e-06	0.995	0.984	0.945	0.909	0.861	0.787
	0.75	1.102139e-05	0.992	0.980	0.953	0.918	0.882	0.774
	0.90	1.322566e-05	0.988	0.959	0.933	0.888	0.835	0.723
	1.00	1.469518e-05	0.992	0.965	0.933	0.900	0.849	0.750
	1.10	1.616470e-05	0.982	0.943	0.906	0.857	0.786	0.695
	1.25	1.836898e-05	0.990	0.953	0.903	0.813	0.772	0.653
	2	2.939036e-05	0.982	0.920	0.819	0.766	0.705	0.593
	4	5.878073e-05	0.965	0.862	0.741	0.666	0.610	0.439
	10	1.469518e-04	0.829	0.668	0.581	0.485	0.412	0.323
	20	2.939036e-04	0.902	0.783	0.722	0.655	0.607	0.520
	100	1.469518e-03	0.944	0.852	0.784	0.718	0.666	0.550
0.112500	0.01	3.076687e-07	0.997	0.978	0.926	0.885	0.850	0.744
	0.10	3.076687e-06	0.997	0.961	0.926	0.885	0.845	0.732
	0.25	7.691717e-06	1.000	0.967	0.923	0.882	0.847	0.754
	0.50	1.538343e-05	0.997	0.971	0.927	0.891	0.849	0.731
	0.75	2.307515e-05	0.997	0.964	0.925	0.884	0.844	0.740
	0.90	2.769018e-05	0.994	0.966	0.921	0.884	0.849	0.741
	1.00	3.076687e-05	0.996	0.954	0.911	0.871	0.823	0.738
	1.10	3.384355e-05	0.997	0.960	0.911	0.878	0.809	0.725
	1.25	3.845858e-05	0.996	0.949	0.904	0.861	0.792	0.715
	2	6.153373e-05	0.992	0.951	0.894	0.850	0.806	0.714
	4	1.230675e-04	0.987	0.933	0.866	0.805	0.748	0.657
	10	3.076687e-04	0.884	0.729	0.634	0.555	0.499	0.418
	20	6.153373e-04	0.843	0.773	0.679	0.604	0.529	0.462
	100	3.076687e-03	0.909	0.823	0.760	0.673	0.626	0.525

Table 2.12: Confidence interval coverage for density at different bandwidths $c \times \hat{a}$ and at different densities ρ_n, based on Monte Carlo simulations using the horseshoe generating function when $n = 500$ and $S = 1000$.

			Proportion of bootstrap CI that cover truth for					
ρ_n	c	average $\hat{a}$	α=0.01	α=0.05	α=0.1	α=0.15	α=0.2	α=0.3
0.37950	0.001	9.252546e-08	0.000	0.000	0.000	0.000	0.000	0.000
	0.01	9.252546e-07	0.000	0.000	0.000	0.000	0.000	0.000
	0.1	9.257237e-06	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	2.314309e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.5	4.628619e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.75	6.942928e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.9	8.331514e-05	0.000	0.000	0.000	0.000	0.000	0.000
	1	9.257237e-05	0.000	0.000	0.000	0.000	0.000	0.000
	1.1	1.018296e-04	0.000	0.000	0.000	0.000	0.000	0.000
	1.25	1.157155e-04	0.000	0.000	0.000	0.000	0.000	0.000
	2	1.851447e-04	1.000	0.990	0.983	0.956	0.920	0.852
	4	3.702895e-04	0.999	0.993	0.981	0.953	0.897	0.820
	10	9.257237e-04	0.999	0.987	0.941	0.896	0.845	0.765
	100	9.252546e-03	0.997	0.959	0.899	0.828	0.773	0.661
	1000	9.252546e-02	0.997	0.955	0.899	0.833	0.791	0.668
0.56925	0.001	1.108253e-07	0.000	0.000	0.000	0.000	0.000	0.000
	0.01	1.108253e-06	0.000	0.000	0.000	0.000	0.000	0.000
	0.1	1.107591e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	2.768978e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.5	5.537956e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.75	8.306935e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.9	9.968322e-05	0.000	0.000	0.000	0.000	0.000	0.000
	1	1.107591e-04	0.000	0.000	0.000	0.000	0.000	0.000
	1.1	1.218350e-04	0.000	0.000	0.000	0.000	0.000	0.000
	1.25	1.384489e-04	0.000	0.000	0.000	0.000	0.000	0.000
	2	2.215183e-04	1.000	0.990	0.974	0.956	0.936	0.853
	4	4.430365e-04	0.998	0.976	0.938	0.889	0.837	0.744
	10	1.107591e-03	0.983	0.931	0.842	0.771	0.705	0.585
	100	1.108253e-02	0.907	0.738	0.505	0.435	0.346	0.259
	1000	1.108253e-01	0.885	0.640	0.495	0.437	0.380	0.259
0.75900	0.001	9.286943e-08	0.000	0.000	0.000	0.000	0.000	0.000
	0.01	9.286943e-07	0.000	0.000	0.000	0.000	0.000	0.000
	0.1	9.277168e-06	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	2.319292e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.5	4.638584e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.75	6.957876e-05	0.000	0.000	0.000	0.000	0.000	0.000
	0.9	8.349452e-05	0.000	0.000	0.000	0.000	0.000	0.000
	1	9.277168e-05	0.000	0.000	0.000	0.000	0.000	0.000
	1.1	1.020489e-04	0.030	0.001	0.000	0.000	0.000	0.000
	1.25	1.159646e-04	0.860	0.650	0.449	0.303	0.221	0.133
	2	1.855434e-04	0.999	0.982	0.954	0.926	0.879	0.787
	4	3.710867e-04	1.000	0.981	0.952	0.910	0.856	0.755
	10	9.277168e-04	0.982	0.914	0.867	0.793	0.735	0.650
	100	9.286943e-03	0.929	0.865	0.738	0.648	0.595	0.479
	1000	9.286943e-02	0.906	0.772	0.659	0.605	0.535	0.444

Table 2.13: Confidence interval coverage for λ_{10} at different bandwidths $c \times \hat{a}$ and at different densities ρ_n, based on Monte Carlo simulations using the high density generating function when $n = 500$ and $S = 1000$.

Subsection 2.B.3 Plots

The linking functions we consider are:

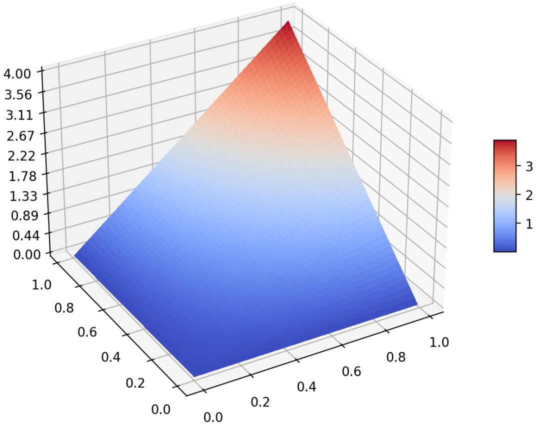


Figure 2.12: Dot product linking function: $h(\xi_i, \xi_j) = \rho_n \times 4\xi_i\xi_j$.

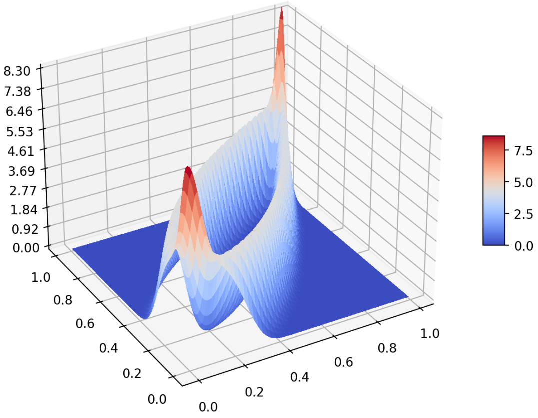


Figure 2.13: Horseshoe linking function:
$h(\xi_i, \xi_j) = \rho_n \times 4.44286 \left(e^{-200\left(\xi_i - \xi_j^2\right)^2} + e^{-200\left(\xi_j - \xi_i^2\right)^2} \right)$.

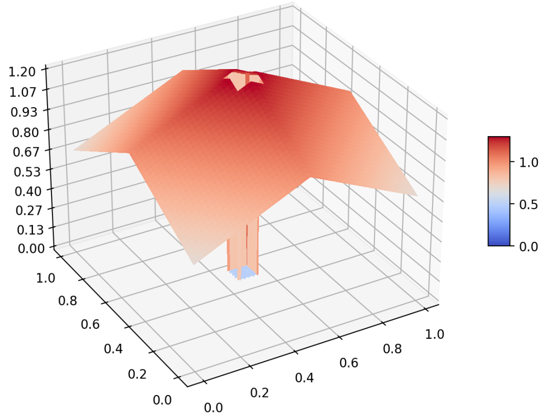


Figure 2.14: High-density linking function:
$h(\xi_i, \xi_j) = \rho_n \times 1.35 \left(1 - \mathbb{1}\left(\left|\frac{1}{2} - \xi_i\right| \leq 0.05\right) \mathbb{1}\left(\left|\frac{1}{2} - \xi_j\right| \leq 0.05\right)\right) \left(1 - \frac{1}{2}\left(\left|\frac{1}{2} - \xi_i\right| + \left|\frac{1}{2} - \xi_j\right|\right)\right)$.

Appendix 2.C Extensions and alternative specifications

Subsection 2.C.1 Possible extensions of application

On a technical level, one possible extension would be analysing different ways of aggregating the twelve observed types of household interactions into the adjacency matrix. Like the original paper, we have used the union of the twelve characteristic-specific adjacency matrices, but there are many other possible choices, e.g. taking an intersection (this may be less desirable as it leads to a significantly sparser network) or an average (which gives a weighted adjacency matrix). Our method allows for comparison of the adjacency matrices achieved though different aggregating functions and checking if they lead to different structures. For example, we can compare the largest eigenvalues λ_1 obtained for villages using different aggregating functions and check if they have overlapping confidence intervals. If they do not, this shows that the choice of the aggregating function is not without loss of generality.

Another possible modification of our procedure would be estimating the linking function under the assumption that each of the twelve characteristics is a separate draw from the Bernoulli distribution. We could redefine the distance function to depend directly on the twelve characteristics instead of a single aggregate adjacency matrix. Let $\mathbf{A}_{ij}$ denote a 12×1 vector of indicators whether households i and j are related according to the twelve characteristics. Let $\|.\|$ be some vector norm (e.g. max norm, min norm, a weighted norm,[29] or a Euclidean norm). Then we can define

$$d_{ij}^{(\|\cdot\|,2)} - \left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\sum_{s=1}^{n}\left\|\operatorname{diag}\left(\mathbf{A}_{ts}\right)\left(\mathbf{A}_{is} - \mathbf{A}_{js}\right)\right\|\right)^2\right)^{\frac{1}{2}}$$

[29] A weighted norm could be of the form $\|x\| = \sqrt{x' W_x x}$, with a weight matrix W_x that may depend on the adjacency matrices, e.g. $W_x = \left(\frac{1}{n}\sum_{v=1}^{n} x_v x_v'\right)^{-1}$.

where diag(v) is a diagonal matrix with diagonal entries from a vector v. To obtain the bootstrap version of adjacency matrices $\mathbf{A}$ we could draw from a joint Bernoulli distribution with probabilities estimated using $\hat{h}_n$ based on the distance $d_{ij}^{(\|\cdot\|,2)}$ and the adjacency matrices for individual characteristics, and a covariance matrix equal to the sample covariance between different characteristics.

Subsection 2.C.2 Sensitivity checks in application

We rerun the estimation for a subset of villages using 300 repetitions of the simulated information spreading through the network to estimate the simulated moments instead of the original 75. The outcomes (middle panel in Fig. 2.15) for that subsample were very similar to the outcomes based on the original specification (left panel in Fig. 2.15). We conclude that 75 simulations are sufficient for the estimation of simulated moments.

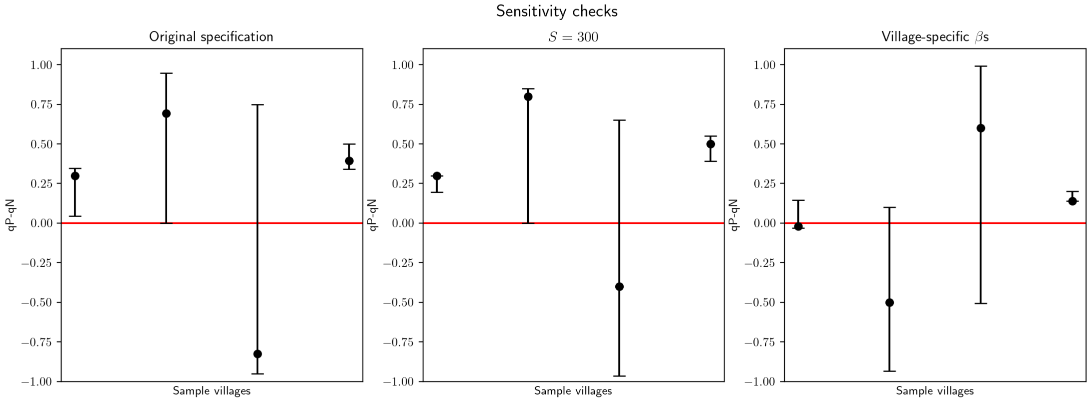


Figure 2.15: A comparison of the estimates of $q^P - q^N$ for a subset of villages with 95% confidence intervals based simulated moments estimated using the original specification with 75 simulations and β estimated using all villages (left), 300 simulations and β estimated using all villages (middle), and 75 simulations and β estimated using village-specific data only (right).

We also tried to estimate the β coefficients using village-specific data (rather than aggregating over all villages). This did make a difference for the estimates and confidence intervals (right panel of Fig. 2.15), though the conclusions remain similar. However, since the regression used to identify β is run using only the information about the leaders, we found the sample sizes for individual villages too small to give reliable estimates. Hence we chose to use aggregate β in our main simulations.

Subsection 2.C.3 Alternative bootstrap procedure: links only

Instead of the procedure we use in the main paper, we could skip the step 2. of resampling nodes from the original graph and go straight into resampling links according to $\hat{h}_n$ for the original set of nodes. The motivation for this procedure is similar to the one we use: the original sample comes from the true data generating distribution, we have a good estimate for the distribution of the adjacency matrix, hence the networks simulated this way should preserve the structure of the original network. Skipping one step in the simulation simplifies the procedure, improves computational time and would also simplify the proofs. We run some simulations using this method and found that, unfortunately, it does not perform as well as our main approach. Table Table 2.14 shows the results of some of our simulations. We see that the confidence interval coverage is very poor, other than for a few special cases where the sample size is small ($n = 25$, in which case the bias is small relative to variance and the true value may still be included in the confidence interval) or we are estimating a statistic which is relatively tricky to estimate (e.g. λ_2) and hence measured with more variation than e.g. λ_1 or density.

			95% CI coverage for			
generating function	n	ρ_n	density	transitivity	λ_1	λ_2
high density	25	0.379500	0.803	0.848	0.791	0.978
		0.569250	0.803	0.788	0.776	0.979
		0.759000	0.755	0.730	0.752	0.973
	100	0.379500	0.541	0.537	0.495	0.824
		0.569250	0.509	0.554	0.528	0.925
		0.759000	0.532	0.518	0.529	0.963
	300	0.379500	0.168	0.144	0.156	0.662
		0.569250	0.211	0.222	0.227	0.824
		0.759000	0.340	0.337	0.339	0.947
	500	0.379500	0.068	0.059	0.058	0.410
		0.569250	0.087	0.110	0.092	0.673
		0.759000	0.251	0.282	0.260	0.950
horseshoe	25	0.056250	0.727	0.920	0.838	0.956
		0.084375	0.776	0.910	0.770	0.936
		0.112500	0.745	0.757	0.715	0.916
	100	0.056250	0.530	0.520	0.406	0.819
		0.084375	0.616	0.419	0.387	0.717
		0.112500	0.686	0.365	0.359	0.679
	200	0.056250	0.430	0.344	0.243	0.659
		0.084375	0.585	0.308	0.302	0.606
		0.112500	0.592	0.270	0.287	0.501
	300	0.056250	0.433	0.285	0.215	0.575
		0.084375	0.589	0.266	0.227	0.504
		0.112500	0.568	0.237	0.240	0.460
	500	0.056250	0.407	0.213	0.158	0.498
		0.084375	0.530	0.215	0.221	0.413
		0.112500	0.490	0.217	0.218	0.326
product	25	0.125000	0.562	0.907	0.652	0.911
		0.250000	0.484	0.785	0.513	0.963
	100	0.125000	0.331	0.567	0.366	0.887
		0.250000	0.263	0.423	0.298	0.957
	300	0.125000	0.203	0.253	0.164	0.867
		0.250000	0.124	0.177	0.136	0.914
	500	0.125000	0.156	0.165	0.140	0.836
		0.250000	0.089	0.126	0.090	0.865

Table 2.14: 95% confidence interval coverage for density, transitivity, λ_1 and λ_2 for different generating functions, different sample sizes n from 25 ot 500 and at different densities ρ_n, based on Monte Carlo simulations using a version of the algorithm which keeps the original set of nodes and only resamples the links between them.

Our hypothesis for why the performance is so poor is that if the bandwidth a_n is small, or if for some individuals there are no close neighbours, our procedure becomes similar to the empirical bootstrap of Green and Shalizi (2022): we draw a link between two individuals if and

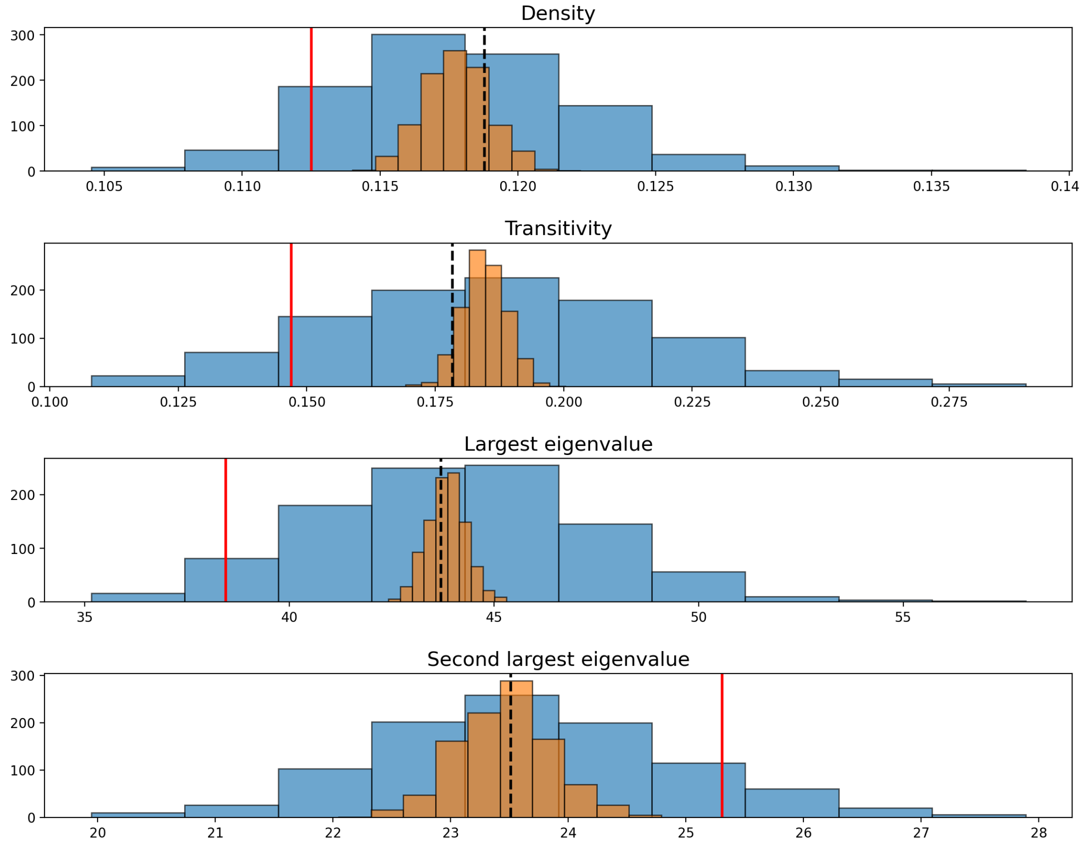


Figure 2.16: Comparison of histograms for our main bootstrap approach (in blue) and for the links-only version (in orange) from estimation of one specific network from the horseshoe generating function with $n = 300$ and $\rho_n = 0.1125$. The estimated statistics are, from top to bottom: density, transitivity, λ_1 and λ_2. The red solid line denotes the population true value of the statistic while the black dashed line denotes the value of the statistic in the bootstrapped graph.

only if they were linked in the original graph. Hence if we do not resample individuals, the bootstrapped graph may become too similar to the original graph, at least on the subgraphs consisting of individuals with few neighbours. This can lead to insufficient variation in the bootstrapped graphs and worse performance of the bootstrap procedure. Fig. 2.16 and Fig. 2.17 shows a comparison of our original method (in blue) and the version which only resamples links while keeping the original nodes (in orange). While both do a good job of replicating the statistic values in the bootstrapped graph (dashed black line), the version which only resamples links is too concentrated around the value in the bootstrapped graph and often misses the population value of the statistic (red solid line), leading to poor confidence interval coverage of the links-only procedure.

This indicates that if one is interested in uncovering the population values, our main procedure is more reliable. However, in applications where we are only interested in confidence intervals for a specific sample, bootstrapping links only does provide narrower confidence intervals and would be preferred.

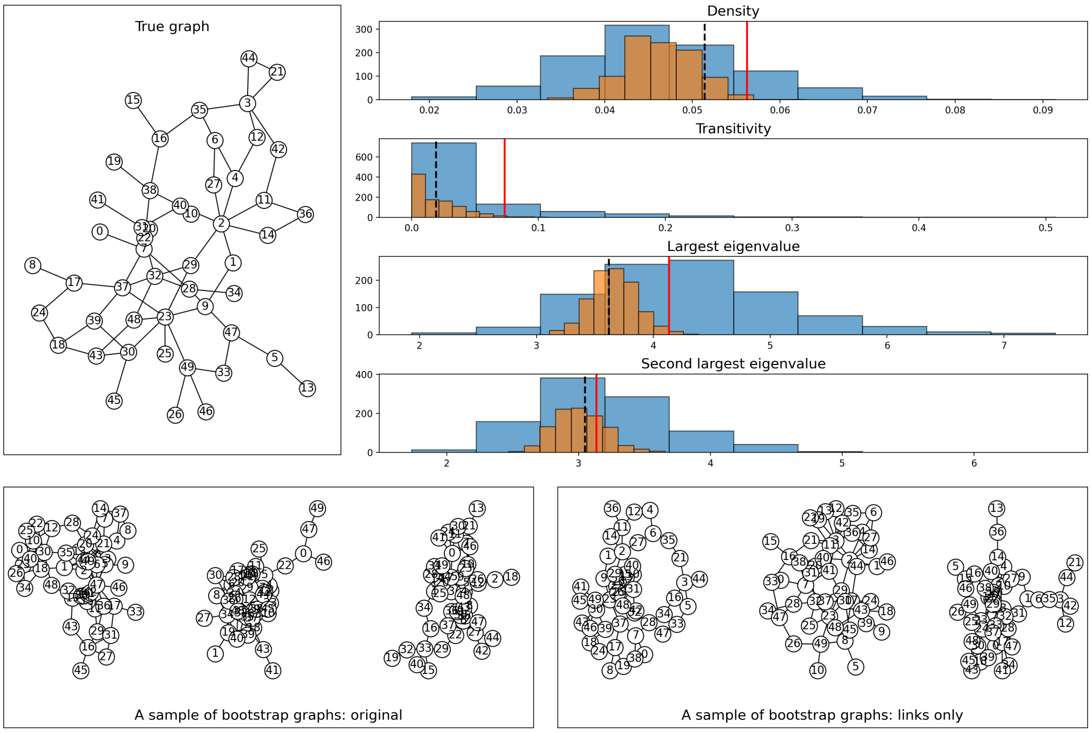


Figure 2.17: Comparison of histograms for our main bootstrap approach (in blue) and for the links-only version (in orange) from estimation of one specific network from the horseshoe generating function with $n = 50$ and $\rho_n = 0.05625$. The estimated statistics are, from top to bottom: density, transitivity, λ_1 and λ_2. The red solid line denotes the population true value of the statistic while the black dashed line denotes the value of the statistic in the bootstrapped graph.

Chapter 3

Understanding regression shape changes through nonparametric testing

Abstract

We propose a procedure for testing whether a nonparametric regression mean satisfies a shape restriction that varies within the domain of the regressor. Notably, the change points of these shape restrictions are unknown and must be estimated. Our test statistic is based on the empirical process, drawing inspiration from Khmaladze (1982). This paper extends the nonparametric methodology of Komarova and Hidalgo (2023) by proposing a method to estimate the shape change points and consequently addressing the additional estimation errors introduced by that stage. We analyse strategies for managing these errors and adapting the testing approach accordingly. Our framework accommodates various common shapes, such as (inverse) U-shapes, S-shapes, and W-shapes. Furthermore, our method is applicable to partial linear models, thereby encompassing a broad spectrum of applications. We demonstrate the efficacy of our approach through application to several economic problems and data.

3.1 Introduction

This paper proposes a nonparametric procedure for testing shape constraints of the regression mean when the shape changes within the domain of the regressor. We allow the shape to potentially change not just once but multiple times. In economics and other disciplines changing shape patterns are encountered frequently. There is seizable empirical literature analysing or attempting to establish U-shaped or hump-shaped relations. S-shape relations are also frequently encountered in economics in the context of poverty traps or innovations. If U-shapes can be described as changing a shape pattern only once in the domain – from decreasing to increasing, the S-shapes in a certain interpretation may involve a more complicated characterisation as there are not just three different monotonicity patterns (down, then up, then down again) but also a switch from convex to concave.

A common empirical practice in cases of U-shapes or hump-shapes has often relied on quadratics despite some recognising a potential need for nonparametric approaches[1]. The use of quadratic specifications, however, remains a widespread practice. It is intuitive why – quadratics may be appealing to researchers due to the simplicity of interpretation and ability to model both convex and concave responses. It is evident, however, that just using quadratics is a restrictive way to model nonlinearities, as: (a) it imposes symmetry around the turning point; (b) it has to be concave (convex) everywhere; (c) if it is concave (convex), it first has to decrease (increase) and then increase (decrease). In reality, nonlinear relationships may be much more complicated than that. We refer the reader to Section 3.A, where we present examples of nonlinear relationship which will be either completely missed or largely misrepresented by the use of quadratics. This will be further confirmed and illustrated in our applications. We believe that the main reason why applied researchers have relied extensively on quadratic specifications has been the absence of a unifying nonparametric methodology that can be used for estimation and testing of nonlinear shapes in a variety of models, including partially linear models. This is precisely where our paper makes its contribution.

Some important and welcome developments in the context of U-shaped or inverse U-shaped relations have been made in the work of Lind and Mehlum (2010), Simonsohn (2018), Kostyshak (2017) and Ganz (2024). These works are discussed in more detail below. Komarova and Hidalgo (2023) discuss U-shapes and S-shapes as some of the applications of their method but they rely on the shape changing point(s) to be known which may be unrealistic in practice. Even though this paper builds to a certain extent on the methodology in Komarova and Hidalgo (2023), it makes an important step forward in allowing the turning points to be unknown. This

[1] For example, nonparametric fits for some of their specification were explored in Aghion, Van Reenen, and Zingales (2013) through their Lowess smoother in Figure 1, or Ashraf and Galor (2013) nonparametric fit in Figure 3, or Aghion et al. (2005) spline fit in Figure II.

implies the needs to estimate them and incorporate the extra estimation steps into the testing methodology, which is a nutshell description of the contributions of our paper. Our extension of the method in Komarova and Hidalgo (2023) is far from trivial. Theoretically, we have to make sure that the turning points are estimated in a way that their interference with the statistical properties of the test statistics is limited in a sample and is also asymptotically negligible. Practically, allowing the turning points to be found adaptively from the data adds robustness to the hypothesis testing and credibility to the conclusions from the test. It is worth emphasising that our testing method applies to a very general class of regression function changing shapes which goes beyond just U-shapes and S-shapes. Additionally, our paper makes it explicit how to apply the testing methodology to partially linear model where the shape constraints enter through the nonparametric part and the effect of other variables is allowed through a linear index.

The paper proceeds as follows. Section 3.2 gives a literature review. Section 3.3 discusses the basic setting and gives a brief overview of our testing methodology. Section 3.4 gives details of the estimation procedure and the test specification. Section 3.5 describes our testing approach in detail. Section 3.6 presents Monte Carlo simulations. Section 3.7 contains applications. Section 3.8 concludes. Section 3.A describes additional motivating examples and Section 3.B.1 includes proofs.

3.2 Literature review

Even though there is no general approach in the literature to estimate and test regression shape changes for various shapes, there is some literature that attempts to address some special cases of this, such as U-shapes and hump-shapes.

The existing literature on testing U-shape constraints is relatively small. Although the shape appears in many settings in economics and social sciences, researchers usually use tests based on quadratic specification. Lind and Mehlum (2010), Simonsohn (2018) and Kostyshak (2017) all give compelling arguments for why tests based on quadratic approximations are not appropriate when testing for a U-shaped or hump-shaped relationship.

Lind and Mehlum (2010) was the first to explicitly highlight the problems with using "U-shaped" and "quadratic" as synonyms. They propose a joint inequality test on the signs of first derivatives estimated at two points in the support. It is a parametric test and it relies on knowing the true functional form. As pointed out by Simonsohn (2018), that test is only valid if the correct functional form is used and is likely to suffer from a high false-positive rate when the model is misspecified.

Simonsohn (2018) proposes a simple test based on estimating two regression lines: for low and high values. He does not assume any functional form but instead tests if the average slope on either side of a switch point is significant, and if the slopes have opposite signs. The switch point is estimated from the data and is chosen to maximise the power of the test (instead of getting the best fit for the data like in our paper) using what the paper calls a "Robin Hood" algorithm, "for it takes away observations from the more powerful line and assigns them to the less powerful one". This test is simple to use but it does have some drawbacks: it does not distinguish between single and multiple changes in the sign of derivative (would classify a W-shape or an N-shape as a U-shape) and its asymptotic properties have not been analysed. In particular, the implication of estimating the switch point to maximise power rather than fit the data are not clear (potentially, this may result in high Type I error).

Kostyshak (2017) uses a non-parametric test, where the test statistic is the smallest bandwidth such that a local polynomial regression is quasi-convex (i.e. U-shaped or monotone), followed by a test for monotonicity. This specification allows for the switch point to be unknown and for the presence of covariates, just like in our model. The test statistic is consistent but further asymptotic theory of the test is not provided. The testing algorithm relies on bootstrap (our test has a nice asymptotic distribution, but to improve the finite sample performance of our test we also resort to bootstrap in this paper). Kostyshak (2017) applies his test to life satisfaction in age and finds that much of the U-shape can be explained by the increase in financial satisfaction typically occurring later in life. A very interesting aspect of that application is that this finding would be completely missed by quadratic specifications. This resonates with our applications too, where we show that a quadratic specification may completely miss a U-shaped or hump-shaped relation. It appears that the idea of the test in Kostyshak (2017) may be extended to other shapes and multiple switch points by relying on more general tests for the number of peaks and valleys in the regression function and its derivatives, but it would require running a series of tests instead of a single test, and would give a researcher less control over the exact choice of a shape than our method.

An approach in a recent work Ganz (2024) is also designed towards testing for U-shape or inverse U-shape relations. The regression function is modelled using linear (first-degree) splines or quadratic I-splines and a candidate switch point is taken as one of the knots (in our approach the switch point is adaptively found first and then the system of knots is driven by the estimated switch point). The idea of Ganz (2024) is to estimate three models – one model is very flexible (not enforcing any constraints), the second one estimates a monotonic relationship, and the third one permits one switch point in line with an (inverse) U-shape relationship. If the fit of the first model is close to that of the third and better than the second, we conclude the relation is (inverse) U-shaped, otherwise we reject the (inverse) U-shaped relation. While the procedure performs well in simulations, the formal statistical properties of this test have not yet been established. It seems that, to ensure a flexible choice of switch point, the number of knots must increase with the sample size, but it is not clear how this affects the asymptotic behaviour of the test.

For more complex changing shapes – those beyond (inverse) U-shapes – to the best of our knowledge, there are no existing statistical testing procedures that allow unknown switch points (Komarova and Hidalgo (2023) can be used when such points are known).

The theoretical and empirical literature has, of course, dealt with non-linear shapes. For examples of U-shaped relationships in economics and other disciplines see e.g.Weiman (1977), Goldin (1995), Calabrese and Baldwin (2001), Groes, Kircher, and Manovskii (2014), Sutton and Trefler (2016) (also see discussions in Lind and Mehlum (2010), Simonsohn (2018) and Kostyshak (2017)). Inverse U-shaped relationships include the case of the so-called single-peaked preferences, which is an important class of preferences in psychology and economics. In empirical research, U-shaped functions e.g. often appear in environmental economics, particularly in studies relating electricity consumption to temperature. Typically, the switch between heating and cooling is set around 18.3°C (65°F). Traditionally, this non-linear relationship between temperature and electricity consumption is modelled using heating degree-days (HDD) and cooling degree-days (CDD) in least squares regressions, as in Pardo, Meneu, and Valor (2002). More advanced techniques, like panel threshold regression (Bessec and Fouquau (2008)) or semiparametric spline models (Engle, Granger, Rice, and Weiss (1986)), have also been used. Another area where U-shaped relationships are found is in the study of happiness across the lifespan. Research by Blanchflower (2020) and others (e.g., Clark (2007)) shows that, after controlling for factors like gender, education, marital status, and employment, happiness follows a U-shape, with a minimum around age 48. This pattern has been observed in both developed and developing countries, and similar findings have been confirmed for apes (Weiss et al. (2012)). However, these studies rely on quadratic specifications in age to test the relationship even when graphical evidence (like in Blanchflower (2020))) a more consistent with an asymmetric U-shape relationship. More advanced techniques, like semiparametric splines (Wunder, Wiencierz, Schwarze, and Küchenhoff (2013)), show a U-shape below age 60 but a downward trend beyond that. In happiness research, identifying the turning point in age is key and, thus, techniques like our would be most suitable also for that reason.

Some literature in accounting documented S-shaped relationships – when e.g. stock price response to unexpected earnings is first convex and then becomes concave after a switch point (and is monotonic throughout the domain). For specific examples see Freeman and Tse (1992) or Das and Lev (1994), among others. S-shaped growth curves of the adopted population in a large society is a generally accepted empirical feature of innovation diffusion (see discussions in Utterback (1996), Rogers (2003)). Thus, testing for an S-shape in this case would allow one to conclude whether technology evolves as one would expect. Newell, Genschel, and Zhang (2014) uses S-shaped curves to model decays in the availability or usage of traditional media. We have not been able to find a formal statistical test in the literature for this type of shape.

An important part of our analysis is estimating switch points between different shape patters (this is formally defined later). There are a few papers using a kernel approximation to estimate a minimum (or maximum) of an unknown function, starting with Parzen (1962) who describes a procedure for finding a mode of a probability density function. Eddy (1980) improves his method to achieve better convergence rate, he shows that the mean squared error of the mode estimator can converge to zero at rate $n^{-1-\varepsilon}$ for any $\varepsilon > 0$. Muller (1989) describes a similar procedure for finding a peak of a regression function. We are not aware of similar procedures using splines.

There is also a large literature on identifying break points in regression functions, i.e. points at which the function is either discontinuous or has a discontinuity in one of its derivatives. For example Feder (1975) develops asymptotic theory for linear estimators of segmented regressions, where the parameters of interest are both the parameters in each segment and points at which the behaviour of the function changes. Estimates of these kinds of break points are typically faster than $n^{-1/2}$ (see e.g. Muller (1992)), making them very attractive, but in this paper we avoid making any assumptions about a level of discontinuity (if any) at the switch points so we don not rely on any result of that kind. Other important papers in this strand of literature on structural breaks include Delgado and Hidalgo (2000), who suggest estimators of location and size of structural breaks in a nonparametric regression model and is applicable in both cross sectional and time series models. Hidalgo, Lee, and Seo (2019) give robust inference in threshold regression models when it is not known a priori whether at the threshold point the true specification has a kink or a jump and the threshold itself is unknown. In a related work, Hidalgo, Lee, Lee, and Seo (2023) propose a continuity test for the threshold regression model based on the findings about a risk lower bound in estimating the threshold parameter without knowing whether the threshold regression model is continuous or not.

Our model also involves a semi-parametric specification combining *B-spline* approximation with linear components, which has been analysed in a number of papers, e.g. Speckman (1988), Heckman (1986). Rice (1986) analyses convergence rates for semiparametric model combining splines and linear terms, under particular assumptions on variables. The main difference between his approach and ours is that he uses an n-dimensional space of splines with all the observations of the regressor treated as knots, whereas we use a space of splines smaller than the number of observations and we can define the knots independently of the data. The basis splines in his case are orthogonal, simplifying derivations, but because of the more dense spline system his model is more prone to overfitting, which he avoids by adding a penalty term. In our case the number of splines grows slower than the number of observations, allowing us to achieve consistency without adding a smoothing penalty term. In his model he shows that the estimate of the linear component is biased, with rate depending on the size of the penalty, and that to decrease bias one needs to use lower penalty than optimal. Under our assumptions we can use the results from Newey (1997) which show that the parametric component achieves root n consistency and is asymptotically unbiased.

3.3 Setting and a brief outline of main ideas

Our leading case in Sections 3.3-3.5 can be described by the following setting:

$$y = m(x) + z'\gamma_0 + u, \tag{3.1}$$

$$E[u|x,z] = 0, \tag{3.2}$$

where $m \in C^1\left[\underline{x}, \overline{x}\right]$ where C^1 denotes the class of smooth functions. The function $m(\cdot)$ and parameter γ_0 is unknown.[2]

To characterise the property of $m(\cdot)$ as that of a *changing shape*, we start with an illustration of a function that changes shape *once*.

Let us, first, denote $m|_{[a,b]}$ as $m(\cdot)$ restricted to the interval $[a,b]$ and, second, suppose that for some $s_1^0 \in [\underline{x}, \overline{x}]$,

$$m|_{[\underline{x}, s_1^0]} \in \mathcal{M}_1\left([\underline{x}, s_1^0]\right), \quad m|_{[s_1^0, \overline{x}]} \in \mathcal{M}_2\left([s_1^0, \overline{x}]\right), \tag{3.3}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are two classes of functions that describe functional properties that can be localised in the sense that

$$m|_{[a,b]} \in \mathcal{M}_j\left([a,b]\right) \quad \Rightarrow \quad m|_{[c,d]} \in \mathcal{M}_j\left([c,d]\right) \quad \forall [c,d] \subseteq [a,b], \quad j = 1, 2. \tag{3.4}$$

We also assume that

$$\mathcal{M}_1\left([a,b]\right) \cap \mathcal{M}_2\left([a,b]\right) = \emptyset \quad \forall [a,b]. \tag{3.5}$$

[2] This setting can be easily extended to allow for $\psi(z, \gamma_0)$ instead of $z'\gamma_0$ for a known nonlinear function $\psi(z, \cdot)$ and unknown γ_0.

We interpret s_1^0 as the *turning* or *switch* point as at that point the regression function changes its pattern from class $\mathcal{M}_1$ to class $\mathcal{M}_2$. We are ultimately interested in a scenario where s_1^0 is not known and has to be estimated from the data.

Consider the following two examples.

Example 1 (U-shape, inverse U-shape, quasi-convexity, quasi-concavity)**.** *To the best of our knowledge, there is no general agreement in the literature on how to define U-shaped relationships mathematically. On of the most common definitions is that the function first decreases till some switch point and then increases. However, some authors would also incorporate convexity requirements into this property. To avoid any ambiguity, below we state explicitly what we mean by U-shape. Our testing procedures can, of course, allow additional convexity requirements.*

A (strict) U-shaped function $m(\cdot)$ is first (strictly) decreasing on some interval $[\underline{x}, s_1^0]$ and then on $[s_1^0, \overline{x}]$ it is (strictly) increasing. Clearly then

$$\mathcal{M}_1\left([a,b]\right) = \left\{m|_{[a,b]} : m'(x) < 0 \text{ a.e. on } [a,b]\right\},$$
$$\mathcal{M}_2\left([a,b]\right) = \left\{m|_{[a,b]} : m'(x) > 0 \text{ a.e. on } [a,b]\right\}.$$

It is easy to see that $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy conditions (3.4) and (3.5) above. In the case of an inverse U-shape (also often called hump-shape), the roles of $\mathcal{M}_1$ and $\mathcal{M}_2$ are reversed.

A non-strict version of U-shape may involve intervals of constancy and can be formulated as non-strict inequalities on the signs of the derivatives.

Related to U-shape is the class of quasi-convex functions which is defined as

$$\Big\{m(\cdot) : \ \forall x_1, x_2 \in [a,b] \ \ \forall \lambda \in [0,1] \ \ m(\lambda x_1 + (1-\lambda)x_2) \le \max\left\{m(x_1), m(x_2)\right\}\Big\}.$$

Function m is quasi-concave if and only if $-m$ is quasi-convex. A smooth function is quasi-convex (-concave) if and only if it first decreases (increases) up to some point and then increases (decreases) incorporating a special case of monotonicity when a switch point is located at one of the boundary points of the interval. For quasi-convex (-concave) functions this switch point may not be known a priori, and thus, it would have to be estimated. This description can be changed to a strict version.

When considering a U-shape property a researcher may want to make further restriction on the function being convex. It is easy to do by adding an inequality $m''(x) > 0$ to the definition of classes $\mathcal{M}_1$ and $\mathcal{M}_2$.

Example 2 (S-shape)**.** *There is no generally agreed on definition of S-shape. E.g. one interpretation defines a (strict) S-shaped as $m(\cdot)$ which is first (strictly) convex and increasing on* *some interval* $[\underline{x}, s_1^0]$ *and then on* $[s_1^0, \overline{x}]$ *it is (strictly) concave and increasing. In our setting this means*

$$\mathcal{M}_1\left([a,b]\right) = \left\{ m|_{[a,b]} : m''(x) > 0 \text{ and } m'(x) > 0 \text{ a.e. on } [a,b] \right\},$$
$$\mathcal{M}_2\left([a,b]\right) = \left\{ m|_{[a,b]} : m''(x) < 0 \text{ and } m'(x) > 0 \text{ a.e. on } [a,b] \right\},$$

if $m(\cdot)$ *is twice differentiable (if not, convexity and concavity can be formulated without involving the derivatives). It is easy to see that* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *satisfy conditions (3.4) and (3.5) above. This interpretation of S-shape is close to prospect theory in behavioural economics (see Kahneman and Tversky (1979).)*

Other fields may understand S-shape differently. E.g., another way to interpret it would be as the regression function first strictly decreasing then strictly increasing and then strictly decreasing again. This interpretation would require two interiors switch points $s_1^0 < s_2^0$ *and three classes* $\mathcal{M}_1([\underline{x}, s_1^0])$, $\mathcal{M}_2([s_1^0, s_2^0])$ *and* $\mathcal{M}_3([s_2^0, \overline{x}])$ *with*

$$\mathcal{M}_1\left([a,b]\right) = \mathcal{M}_3\left([a,b]\right) = \left\{ m|_{[a,b]} : m'(x) < 0 \text{ a.e. on } [a,b] \right\},$$
$$\mathcal{M}_2\left([a,b]\right) = \left\{ m|_{[a,b]} : m'(x) > 0 \text{ a.e. on } [a,b] \right\}.$$

More generally, we have an *ordered* sequence of interior switch points $s_1^0, s_2^0 \dots, s_J^0$ such as

$$s_0^0 \equiv \underline{x} < s_1^0 < s_2^0 \dots < s_J^0 < \overline{x} \equiv s_{J+1}^0$$

(where the support boundaries are denoted as s_0^0 and s_{J+1}^0 for notational convenience) and a sequence of properties $\mathcal{M}_j$, $j = 1, \dots, J+1$, such that

$$m|_{[s_j^0, s_{j+1}^0]} \in \mathcal{M}_{j+1}\left([s_j^0, s_{j+1}^0]\right), j = 0, \dots, J, \tag{3.6}$$

It is important that the ordering of $\mathcal{M}_j$, $j = 1, \dots, J$ is predetermined – that is, we know the order in which the properties of the regression function change.

Condition C1. *(a) Classes* $\mathcal{M}_j$, $j = 1, \dots, J+1$, *describe functional properties that can be localised in the sense that*

$$m|_{[a,b]} \in \mathcal{M}_j\left([a,b]\right) \quad \Rightarrow \quad m|_{[c,d]} \in \mathcal{M}_j\left([c,d]\right) \quad \forall [c,d] \subseteq [a,b], \quad j = 1, 2. \tag{3.7}$$

(b) We also assume that

$$\mathcal{M}_j([a,b]) \cap \mathcal{M}_{j+1}([a,b]) = \emptyset \quad \forall [a,b], \quad j = 1, \ldots, J. \tag{3.8}$$

Part (a) of Condition C1 refines the notion of what it means for a class to capture *shape* – this is a property that extends to subintervals. Part (b) gives a general condition for a *change in shape* that is formulated for any two consecutive classes.

We can establish the identification of s_j^0, $j = 1, \ldots, J$. Henceforth, $s_1 < s_2 < \ldots < s_J$ will denote a generic ordered sequence of switch points located in the interior of $[\underline{x}, \overline{x}]$.

Proposition 1 (Identification)**.** *In the model (3.6) with a given ordering $s_1^0 < s_2^0 < \ldots < s_J^0$ of switch points, the switch points s_j^0, $j = 1, \ldots, J$, are identified under Condition C1.*

Below is an example of a situation with multiple switch points.

Example 3 (two local regression peaks)**.** *Consider the case when the smooth regression function has two local regression peaks. Then, in addition to estimating the two locations of local regression peaks we have to estimate another point between them where the regression function has a local minimum and turns from the decreasing pattern to the increasing one.*

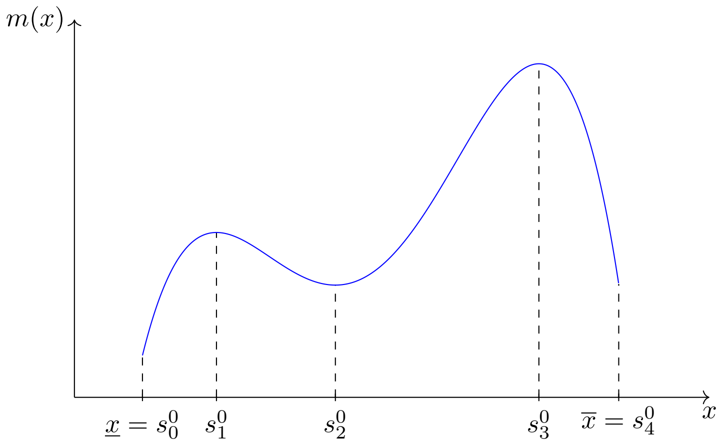


Figure 3.1: Two local regression peaks

Formally we have three interior switch points s_1^0, s_2^0, s_3^0 such that $s_0^0 \equiv \underline{x} < s_1^0 < s_2^0 < s_3^0 < \overline{x} \equiv s_4^0$, with the corresponding sets

$$\mathcal{M}_1([a,b]) = \mathcal{M}_3([a,b]) := \left\{ m|_{[a,b]} : m'(x) > 0 \text{ a.e. on } [a,b] \right\},$$

$$\mathcal{M}_2([a,b]) = \mathcal{M}_4([a,b]) := \left\{ m|_{[a,b]} : m'(x) < 0 \text{ a.e. on } [a,b] \right\}.$$

Points s_1^0 and s_3^0 are locations of the two local regression peaks whereas s_2^0 describes the location of the inevitable local minimum between s_1^0 and s_3^0.

It is easy to see that $\mathcal{M}_j$, $j = 1, \ldots, 4$, satisfy conditions (3.7) and (3.8).

Let $\mathcal{M}_0$ denote the class of all smooth regression functions m that satisfy (3.6):

$$\mathcal{M}_0 = \left\{ m \ : \ m \text{ satisfies (3.6) for some } s_1^0 < s_2^0 < \ldots < s_J^0 \right\} .$$

Our null hypothesis is

$$H_0 : m \in \mathcal{M}_0 \quad vs. \quad H_1 : m \notin \mathcal{M}_0 \tag{3.9}$$

(with the smoothness of functions in $\mathcal{M}_0$ being the maintained hypothesis).

The first step of our testing procedure will be to estimate m by a smooth join of *B-splines* of degree q_j defined on each estimated shape interval $[\hat{s}_j, \hat{s}_{j+1}]$, with $\hat{s}_0 = \underline{x}$, $\hat{s}_{J+1} = \overline{x}$. Suppose that the *B-spline* on $[\hat{s}_j, \hat{s}_{j+1}]$ is build on L_{j+1} *base B-splines* (further details are in the next section). Our estimation will guarantee that as the sample size increases and all L_{j+1}, $j = 0, \ldots, J$, increase with it our estimator will be a consistent estimator of m under H_0.

Before advancing to the detailed technical description of that step, as well as the subsequent steps in testing, let us indicate what will differentiate our method from some other methods available in the literature.

From a big picture perspective, our methodology, just as in Komarova and Hidalgo (2023), is related to methods used in goodness of fit tests. Following Stute (1997) or Andrews (1997) and Komarova and Hidalgo (2023), we base the testing procedure on functionals of the partial sums empirical process

$$\mathcal{K}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i \mathbb{1}(x_i < x), \qquad x \in [\underline{x}, \overline{x}] \tag{3.10}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Here

$$\widehat{u}_i = y_i - \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) - z_i' \widehat{\gamma}, \quad i = 1, \ldots, n,$$

are the residuals obtained after m has been estimated by the nonparametric estimator $\widehat{m}_{\mathcal{B}}(x_i; \hat{s})$ by means of *B-splines* briefly described above and γ_0 has been estimated by $\widehat{\gamma}$ found simultaneously with $\widehat{m}_{\mathcal{B}}(x_i; \hat{s})$, see Section 3.4 for more detail (in a nutshell, $m_{\mathcal{B}}(x; \hat{s}) + z'\widehat{\gamma}$ denotes the best approximation of $m(x) + z'\gamma_0$ using the sum of the join of *B-splines* based on estimated switch points $\hat{s}$ for m and an additive separable linear function in z.) .

Unfortunately, after normalisation, the limit covariance structure of $\mathcal{K}_n(x)$ depends on $\mathcal{M}_0$, making inferences based on $\mathcal{K}_n(x)$ very difficult to perform, if at all possible. For the simplicity of an illustration, consider the case of having no z_i on the right-hand side. Then

$$\mathcal{K}_n(x) = \frac{1}{n}\sum_{i=1}^{n} u_i \mathbb{1}(x_i < x) + \frac{1}{n}\sum_{i=1}^{n}\left(m_{\mathcal{B}}(x_i;\hat{s}) - \widehat{m}_{\mathcal{B}}(x_i;\hat{s})\right)\mathbb{1}(x_i < x) + \frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right)\mathbb{1}(x_i < x). \quad (3.11)$$

In this decomposition the first term can be shown to be $\sqrt{n}$-convergent in distribution to the standard Brownian motion. The second term is also $O_p\left(\frac{1}{\sqrt{n}}\right)$, which means that the asymptotic distribution of $\mathcal{K}_n(x)$ might not be Gaussian and is difficult to characterise, making inferences very cumbersome. The third term in its turn can be represented as

$$\frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right)\mathbb{1}(x_i < x) = \frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - m_{\mathcal{B}}(x_i;s^0)\right)\mathbb{1}(x_i < x) + \frac{1}{n}\sum_{i=1}^{n}\left(m_{\mathcal{B}}(x_i;s^0) - m_{\mathcal{B}}(x_i;\hat{s})\right)\mathbb{1}(x_i < x), \quad (3.12)$$

where the asymptotic behaviour of first sub-term can be made asymptotically negligible with the choice of rates of L_j, $j = 1, \ldots, J$, relative to n and the asymptotic behaviour of the second sub-term depends on the rate of convergence of $\hat{s}$ to s^0 and may be of order $O_p\left(\frac{1}{\sqrt{n}}\right)$, same order of magnitude as the leading term and non-pivotal. When we do include a term linear in z, it creates another non-trivial $O_p\left(\frac{1}{\sqrt{n}}\right)$ component in the test statistic.

In contrast to our method, the approach in Komarova and Hidalgo (2023) would assume the turning points in s^0 to be known (effectively making $\hat{s} = s^0$ in the decomposition above) and, thus, the right-hand side of (3.12) would only have the first sub-term significantly simplifying the ability to control the asymptotic behaviour of that whole term. Our setting is more realistic as the turning points s^0 are taken to be *unknown*. This is a fundamental difference between this paper and Komarova and Hidalgo (2023) which results in very non-trivial theoretical and empirical challenges.

With estimates $\hat{s}$ and $\widehat{m}_{\mathcal{B}}(x_i;\hat{s})$ in hand we apply the transformation of $\mathcal{K}_n(x)$ similar to the one used in Komarova and Hidalgo (2023) and based on ideas of Khmaladze (1982) as well as related to the *CUSUM* of recursive residuals proposed by Brown, Durbin, and Evans (1975). This leads to the asymptotic behaviour of the transformation to be $\sqrt{n}$-convergent to a standard Brownian motion. Then testing is implemented using standard functionals such as Kolmogorov-Smirnov, Cramér -von-Mises or Anderson-Darling. In the next section we give the details of the estimation and testing procedure.

3.4 Modified null hypothesis and estimation methodology

We start with the discussion of estimating m under the null in line with our outline in Section 3.3.

For a given collection of switch points in the vector s, we can consider individual intervals $[s_{j-1}, s_j]$. On each of these intervals we consider a *B-spline* of degree q_j with knots that split $[s_{j-1}, s_j]$ into L'_j equally spaced intervals:[3]

$$m_{\mathcal{B};j}(x;s) \equiv \sum_{\ell=1}^{L_j} \beta_{\ell,j} p_{\ell,L_j,[s_{j-1},s_j],q_j}(x), \quad \text{where } L_j = L'_j + q_j, \tag{3.13}$$

and $\left\{p_{\ell,L_j,[s_{j-1},s_j],q_j}(\cdot)\right\}_{\ell=1}^{L_j}$ is the collection of the base *B-splines* base for the chosen system of knots and the chosen degree q_j (will be described shortly).

Then we can define

$$m_{\mathcal{B}}(x;s) = \sum_{j=1}^{J} m_{\mathcal{B};j}(x;s) \cdot \mathbb{1}[s_{j-1}, s_j) + m_{\mathcal{B};J+1}(x) \cdot \mathbb{1}[s_J, s_{J+1}], \quad x \in [\underline{x}, \overline{x}]. \tag{3.14}$$

Now we want to delve in more detail in the properties of *B-splines* in (3.13). These *B-splines* are constructed from polynomial pieces joined at some specific points called knots. In (3.14) we use *B-splines* whose domain and the system of knots differ on different sides of switch points. Generally, let q be the degree of a spline, L' be the number of subintervals of $[\underline{s}, \overline{s}]$ on which we define the spline (i.e. the number of polynomial pieces), then $L = L' + q$ is the number of *B-splines* in the basis.

We define the system of knots which split $[\underline{s}, \overline{s}]$ into L' equally spaced intervals. When defining *B-spline* of degree q we repeat the knots at the end points of the domain $q+1$ times. To be precise, we let

$$t = (t_\ell)_{\ell=1}^{L+2q+1} = \left(\underbrace{\underline{s}, \ldots, \underline{s}}_{q+1 \text{ times}}, \underline{s} + \frac{\overline{s} - \underline{s}}{L'}, \underline{s} + 2\frac{\overline{s} - \underline{s}}{L'}, \ldots, \underbrace{\overline{s}, \ldots, \overline{s}}_{q+1 \text{ times}}\right).$$

be the knot sequence. Then the ℓth *B-spline* of degree q defined on the knots t is a function of x we denote by $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$. *B-splines* are defined recursively (see De Boor 1978) as follows:

$$p_{\ell-q,L-q,[\underline{s},\overline{s}],0}(x) = \mathbb{1}\left(x \in [t_\ell, t_{\ell+1})\right) = \begin{cases} 1 & \text{if } t_\ell \le x < t_{\ell+1} \\ 0 & \text{otherwise} \end{cases}$$

[3] The condition that these intervals are equally spaced is not important and is only imposed for the simplicity of the exposition. We only need that the system of knots has to become increasingly dense in $[s_{j-1}, s_j]$.

and for $0 < k \leq q-1$:

$$p_{\ell,L-k,[\underline{s},\overline{s}],q-k} = \frac{x-t_\ell}{t_{\ell+q}-t_\ell} p_{\ell-1,L-k-1,[\underline{s},\overline{s}],q-k-1}(x) + \frac{t_{\ell+q+1}-x}{t_{\ell+q+1}-t_{\ell+1}} p_{\ell,L-k-1,[\underline{s},\overline{s}],q-k-1}(x).$$

By convention, anything divided by zero is zero.

An example of the steps in the construction of base *B-splines* for $q=3$, $L=8$, $[\underline{s},\overline{s}]=[0,1]$ is given in Figure 3.2.

Below is the list of some properties of *base B-splines*.

- $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$ is non-negative and is positive over a domain spanned by $q+2$ adjacent knots, and is zero everywhere else;
- between each pair of consecutive knots $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$ is a polynomial of degree q;
- at a knot which is repeated m times $p_{\ell,L,[\underline{s},\overline{s}],q}(x)$ has $q-m$ continuous derivatives;
- at any given x, at most $q+1$ *B-splines* are non-zero;
- at any given x, the values of all *B-splines* sum to 1: $\forall x \in [\underline{s},\overline{s}] \quad \sum_{\ell=1}^{L} p_{\ell,L,[\underline{s},\overline{s}],q}(x) = 1$.

The derivative of a *B-spline* is composed of polynomial sections of degree $q-1$ defined over the same set of knots (with boundary knots having one less multiplicity), and is itself a *B-spline* of degree one lower. In particular, one can show, e.g. by induction (see e.g. De Boor (1978) or Procházková (2005)), that for a base *B-spline*,

$$\frac{\partial p_{\ell,L,[s_{j-1},s_j],q}}{\partial x} = \frac{q}{t_{\ell+q}-t_\ell} p_{\ell-1,L-1,[s_{j-1},s_j],q-1}(x) - \frac{q}{t_{\ell+q+1}-t_{\ell+1}} p_{\ell,L-1,[s_{j-1},s_j],q-1}(x), \tag{3.15}$$

which means that the derivative of the spline $m_{\mathcal{B};j}(x;s) \equiv \sum_{\ell=1}^{L} \beta_\ell p_{\ell,L,[s_{j-1},s_j],q}(x)$ is

$$\frac{\partial m_{\mathcal{B};j}(x;s)}{\partial x} = q \sum_{\ell=2}^{L} \frac{\Delta\beta_\ell}{t_{\ell+q}-t_\ell} p_{\ell-1,L-1,[s_{j-1},s_j],q-1}(x). \tag{3.16}$$

Note that in the final expression the knots t are still based on the original q, not $q-1$.

Approximation and estimation of the regression mean $m(\cdot)$ by *B-splines* are appealing due to a convenient way to capture shape properties of interest, particularly those based on the derivatives of the regression function (such as U-shape, S-shape, etc.). In other words, the use of *B-splines* helps us to write the class $\mathcal{M}_0$ in (3.9) in terms of restrictions on the coefficients of the *base B-splines* in an approximation to $m(\cdot)$ (this requirement captured formally in Condition C2 below). With $L_j \to \infty$ as $n \to \infty$, $j = 1, \ldots, J+1$, the number of coefficients of the *B-splines* and the number of constraints will increase to infinity.

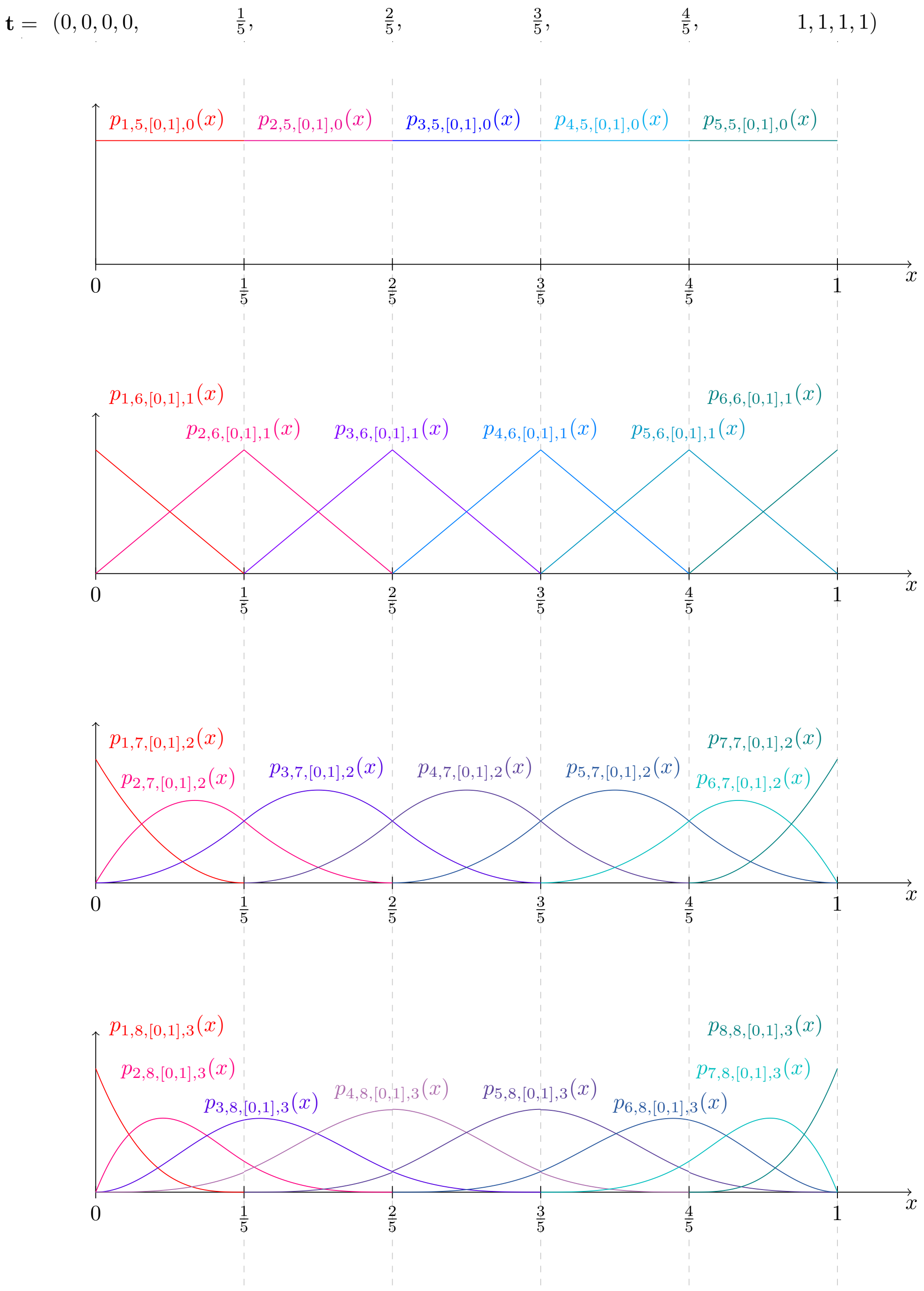


Figure 3.2: An example of *base B-spline* functions construction for $q = 3$.

It is well understood that the choice of the number of knots determines the trade-off between overfitting and underfitting when there are respectively too many or too few knots. The main difference between *B-splines* and *P-splines* is that the latter tend to employ a large number of knots, but to avoid oversmoothing they incorporate a penalty function based on the τ-th difference $\triangle^{\tau}\beta_{\ell}$, where $\triangle\beta_{\ell} = \beta_{\ell} - \beta_{\ell-1}$, with $\tau = 2$ being the most common choice. It is worth mentioning that other sieve estimators might be used, see the survey in Chen (2007), but we found *B-splines* particularly useful for our purposes.

Since our ultimate goal is to develop a nonparametric statistical test for (3.9) using the consistent estimators $\widehat{s}_1, \widehat{s}_2 \ldots, \widehat{s}_J$, we want to be sure that functional properties in each class $\mathcal{M}_j$, $j = 1, \ldots, J{+}1$, can be captured by the properties of coefficients of *B-splines* approximating m on the respective interval $[s_{j-1}, s_j]$ in the partition of $[\underline{x}, \overline{x}]$, and that this representation by the properties of coefficients of approximating *B-splines* becomes both necessary and sufficient as the number of knots on $[s_{j-1}, s_j]$ goes to infinity.

Formally, this is stated in Condition C2 below. Before we formally introduce this condition, let us introduce some helpful notations. Let $\mathcal{B}_j(q_j, L_j)$ denote the set of all *B-splines* of degree q_j with knots that split $[s_{j-1}, s_j]$ into L_j' equally spaced intervals. A generic element in this set is written as a linear combination in (3.13). Thus, any element in $\mathcal{B}_j(q_j, L_j)$ can be fully characterised by the vector $\beta_{all,j} \equiv (\beta_{1,j}, \ldots, \beta_{L_j,j})' \in \mathbb{R}^{L_j}$ and constraints on this vector can be mapped into constraints on the *B-spline*. We consider each vector $\beta_{all,j} \in \mathbb{R}^{L_j}$ to be embedded into the long vector $\beta_{all} = (\beta_{all,1}', \ldots, \beta_{all,J+1}')' \in \mathbb{R}^{\sum_{j=1}^{J+1} L_j}$.

Let $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \subset \mathbb{R}^{\sum_{j=1}^{J+1} L_j}$ denote a set that describes constraints on the vector of coefficients β_{all} for a given vector s pf ordered switch points. We can subsequently define

$$\mathcal{M}_{\{(q_j,L_j)\}_{j=1}^{J+1},s} = \Big\{ m_{\mathcal{B}}(x;s) \text{ in the form of (3.14)} \mid \beta_{all} \in T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \Big\}.$$

$\mathcal{M}_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ is, thus, a collection of functions that are joins of *B-splines* defined individually on the intervals $[s_{j-1}, s_j]$. $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ can contain restrictions that will guarantee that the whole $m_{\mathcal{B}}(\cdot; s)$ defined in such a piece-wise way is continuous, or, additionally, smooth or, more generally, r times continuously differentiable (the choice of r would depend on the degrees q_j, $j = 1, \ldots, J{+}1$ of the *B-splines*). E.g., the continuity of the whole piece-wise approximation is ensured by the constraints

$$\beta_{L_j;j} = \beta_{1;j+1}, \quad j = 1, \ldots, J. \tag{3.17}$$

In order to guarantee the smoothness of the approximation $m_{\mathcal{B}}(\cdot; s)$, in addition to (3.17) we have to impose that[4]

$$\frac{q_j L_j' \left(\beta_{L_j;j} - \beta_{L_j-1;j}\right)}{s_j - s_{j-1}} = \frac{q_{j+1} L_{j+1}' \left(\beta_{2;j+1} - \beta_{1;j+1}\right)}{s_{j+1} - s_j}, \quad j = 1, \ldots, J, \tag{3.18}$$

which simplifies to

$$\beta_{L_j;j} - \beta_{L_j-1;j} = \beta_{2;j+1} - \beta_{1;j+1} = 0, \quad j = 1, \ldots, J \tag{3.19}$$

in the case when the switch point is a local minimum or a local maximum. Further restrictions can be derived to enforce the continuity of the second derivative, etc. shall a researcher want to impose higher order restrictions.

In the regularity conditions in Section 3.5.1 we require the regression function to be r times continuously differentiable, therefore it is natural to narrow down $\mathcal{M}_0$ to include only those regression functions that satisfy this smoothness condition. We will denote this class as $\mathcal{M}_0^*$.

Condition C2 below formalises our idea of approximating the properties in $\mathcal{M}_0^*$ by constraints on coefficients in the approximation $m_{\mathcal{B}}(\cdot; s)$ in a necessary and sufficient fashion.

Condition C2. *For each s there is a set $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \subset \mathbb{R}^{\sum_{j=1}^{J+1} L_j}$ that satisfies the following properties:*

(i) For a given s, $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ does not depend on data $\{x_i\}_{i\in\mathbb{Z}}$ and, thus, is non-stochastic;

(ii) For any s, the boundary of $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ consists of a finite number of smooth surfaces;

(iii) Let $\mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1}}}$ denote the union of $\mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}}$ over all possible s and let $\mathcal{H}$ be the Hausdorff distance in the supremum norm in the space of continuous functions. Then

$$\mathcal{H}\left(\mathcal{M}_0^*, \mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1}}}\right) = O\left(\frac{1}{\left(\min_{j=1,\ldots,J+1} L_j\right)^r}\right) \tag{3.20}$$

for $r > 2$ and $\min_{j=1,\ldots,J+1} L_j \to \infty$.

(iv) Let $f(x_i, s) = m(x_i) - m_{\mathcal{B}}(x_i, s)$ where $m_{\mathcal{B}}(x_i, s)$ is the best possible B-spline *approximation to $m(x_i)$ which satisfies the constraints under H_0^B. If $s \neq s^0$ and x_i is within a neighbourhood of the misspecified switch point in which the incorrect constraint is binding, the $f(x_i, s)$ is proportional to $\|s^0 - s\|$. It follows that if $\mathcal{M}_{0,s^0}^*$ denotes the set of functions in class $\mathcal{M}_0^*$ with switch points s^0. For any pair (s^0, s)*

$$\mathcal{H}\left(\mathcal{M}_{0,s^0}^*, \mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}}\right) = \Omega\left(\|s^0 - s\|_\infty\right). \tag{3.21}$$

[4] This is in case the interior knots are equidistant within each $[s_{j-1}, s_j]$.

Condition C2(i) ensures that the constraints in the estimation can be constructed in a generic fashion and we can talk about a deterministic approximation of class $\mathcal{M}_0^*$ by $\mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}}$. Condition C2$(ii)$ guarantees that for a given s the implementation of conditions $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ comes down to enforcing a finite number of constraints on coefficients in β_{all}. In practice, definitions of $T_{\{(q_j,L_j)\}_{j=1}^{J+1}}$ will often be sufficient to guarantee functional properties of $\mathcal{M}_0^*$. Condition C2(iii) ensures that these conditions become asymptotically necessary. Condition C2(iv) puts a lower bound on the quality of fit when we use a set of constraints which misspecify the switch point: we need the loss in fit to be large enough to allow us to estimate $\hat{s}$. We can think of this condition as saying that, when we use s instead of the true s^0, on the region on which we impose the incorrect constraint, the gap between the best possible fit and the true function is proportional to $\|s-s^0\|_\infty$. Combined Condition C2(iii) and Condition C2(iv) ensure that, as long as $s \to s^0$, the constraints in $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ capture constraints in $\mathcal{M}_{0,s^0}^*$ in a necessary and sufficient way as the number of knots grows to infinity, and if the convergence of s to s^0 is sufficiently fast, the approximation rate in the constrained approximation with the enforced $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ is the same as the rate in the unconstrained *B-spline* approximation. We can interpret r as the number of continuous derivatives elements of $\mathcal{M}_0^*$.

Given Condition C2, our idea is to test the null hypothesis

$$H_0^B: \quad \beta_{all} \in T_{\{(q_j,L_j)\}_{j=1}^{J+1},s} \text{ for some } s \quad vs. \quad H_1^B: \quad \text{(negation of null)} \tag{3.22}$$

formulated in terms of the approximation for m. Note that under the alternative we are still only considering smooth functions within $\mathcal{M}_0^*$.

Let us now illustrate Condition C2 and the approximation for the U-shape in Example 1.

Example 1 (continued). *In the case of the U-shape property the approximation consists of two B-spline joined at s^0:*

$$m_{\mathcal{B}}(x;s^0) = \underbrace{\sum_{\ell_1=1}^{L} \beta_{\ell_1;1} p_{\ell_1,L,[\underline{x},s^0],q}(x)}_{m_{\mathcal{B};1}(x;s^0)} \cdot 1[\underline{x},s^0) + \underbrace{\sum_{\ell_2=1}^{L} \beta_{\ell_2;2} p_{\ell_2,L,[s^0,\overline{x}],q}(x)}_{m_{\mathcal{B};2}(x;s^0)} \cdot 1[s^0,\overline{x}],$$

where for simplicity we took $q_1 = q_2 = q$ (same degree of B-splines on both sides of s^0) and $L_1 = L_2 = L$ (same number of knots on both sides of s^0) and assume $m(\cdot)$ is three times continuously differentiable. To capture monotonicity patterns and also smoothness at s^0 described by (3.17)-

(3.19), we take

$$T_{\{(q,L)\}} = \Big\{ (\beta_{all,1}, \beta_{all,2}) \,|\, \beta_{\ell_1;1} \geq \beta_{\ell_1+1;1},\ \ell_1 = 1, \ldots, L-1,$$
$$\beta_{\ell_2;2} \leq \beta_{\ell_2+1;2},\ \ell_2 = 1, \ldots, L-1,$$
$$\beta_{L;1} = \beta_{L-1;1} = \beta_{1;2} = \beta_{2;2}, \quad \beta_{L_j-2;j} = \beta_{3;j+1}.\}\Big\}$$

Inequalities $\beta_{\ell_1;1} \geq \beta_{\ell_1+1;1}$, $\ell_1 = 1, \ldots, L-1$, capture the fact that the function is decreasing on $[\underline{x}, s^0]$, while $\beta_{\ell_2;2} \leq \beta_{\ell_2+1;2}$, $\ell_2 = 1, \ldots, L-1$, capture the fact that it is increasing on $[\underline{x}, s^0]$. Equality $\beta_{L;1} = \beta_{L-1;1}$ for the continuity of the approximation at s^0, and the equalities $\Delta\beta_{L;1} = \Delta\beta_{2;2} = 0$ for smoothness of the approximation at s^0 as well as for the minimum of the approximation at s^0 together give us $\beta_{L;1} = \beta_{L-1;1} = \beta_{1;2} = \beta_{2;2}$.

Now let us show that C2 (iii) holds. From the B-spline theory we know (e.g. from De Boor (1978)) that the approximation of three times differentiable $m|_{[\underline{x},s^0]}$ and $m|_{[s^0,\overline{x}]}$ by unconstrained B-splines on the respective intervals $[\underline{x}, s^0]$ and $[s^0, \overline{x}]$ can be attained at the rate $O\left(\frac{1}{L^3}\right)$. Let us denote such approximations as $\tilde{m}_{\mathcal{B};1}(\cdot)$ and $\tilde{m}_{\mathcal{B};2}(\cdot)$, respectively:

$$\tilde{m}_{\mathcal{B};1}(\cdot; s^0) = \sum_{\ell_1=1}^{L} \tilde{\beta}_{\ell_1;1} p_{\ell_1,L,[\underline{x},s^0],q}(x), \qquad \tilde{m}_{\mathcal{B};2}(\cdot; s^0) = \sum_{\ell_2=1}^{L} \tilde{\beta}_{\ell_2;2} p_{\ell_2,L,[s^0,\overline{x}],q}(x).$$

Let us show that because of $m|_{[\underline{x},s^0]}$ strictly decreasing we can without a loss of generality take $\tilde{\beta}_{\ell_1;1} \geq \tilde{\beta}_{\ell_1+1;1}$ for all $\ell_1 = 1, \ldots, L-1$, in $\tilde{m}_{\mathcal{B};1}(\cdot)$, and analogously without a loss of generality take $\tilde{\beta}_{\ell_2;2} \leq \tilde{\beta}_{\ell_2+1;2}$ for all $\ell_2 = 1, \ldots, L-1$, in $\tilde{m}_{\mathcal{B};2}(\cdot)$ Indeed, from the approximation theory we know that

$$\sup_{x \in [\underline{x},s^0]} \left| \sum_{\ell_1=1}^{L} \tilde{\beta}_{\ell_1;1} p'_{\ell_1,L,[\underline{x},s^0],q}(x) - m'|_{[\underline{x},s^0]}(x) \right| = O\left(\frac{1}{L^2}\right), \tag{3.23}$$

$$\sup_{x \in [s^0,\overline{x}]} \left| \sum_{\ell_2=1}^{L} \tilde{\beta}_{\ell_2;2} p'_{\ell_2,L,[s^0,\overline{x}],q}(x) - m'|_{[s^0,\overline{x}]}(x) \right| = O\left(\frac{1}{L^2}\right). \tag{3.24}$$

Using the formula for the derivative of B-spline, obtain

$$\sum_{\ell_j=1}^{L} \tilde{\beta}_{\ell_j;j} p'_{\ell_j,L,[s_{j-1},s_j],q}(x) = q \sum_{\ell_j=1}^{L-1} \frac{\triangle\tilde{\beta}_{\ell_j+1'j}}{t^{l_j+1+q;j} - t^{l_j+1;j}} p_{\ell_j+1,L,[s_{j-1},s_j],q-1}(x), \quad j = 1, 2, \tag{3.25}$$

where $t^{l_j;j}$ denotes a knot on $[\underline{x}, s^0]$ for $j = 1$ and on $[s^0, \overline{x}]$ for $j = 2$.

Taking into account (3.23)-(3.25), the fact that $\frac{\underline{K}_1}{L} t^{l_j+1+q;j} - t^{l_j+1;j} \leq \frac{\bar{K}_1}{L}$ for some constant

$\underline{K}_1, \bar{K}_1 > 0$ as well as the facts that $m'|_{[\underline{x},s^0]}(x) \geq 0$ and $m'|_{[s^0,\overline{x}]}(x) \leq 0$ and

$$\sum_{\ell_j=1}^{L} p_{\ell_j,L,[s_{j-1},s_j],q}\,(x) = 1 \quad \textit{for all } x \textit{ in the respective interval,} \tag{3.26}$$

we conclude that

$$\triangle\tilde{\beta}_{\ell_1+1;1} \leq \frac{K_2}{L^3}, \quad \triangle\tilde{\beta}_{\ell_2+1;2} \geq -\frac{K_2}{L^3},$$

for some constant $K_2 > 0$. Thus, to ensure that $\tilde{\beta}_{\ell_1+1;1} \leq 0$, $\ell_1 = 1, \ldots, L-1$, and $\tilde{\beta}_{\ell_2+1;2} \geq 0$, $\ell_2 = 1, \ldots, L-1$, which will guarantee the desired monotonicity patterns in the approximation, we have to change each coefficient $\tilde{\beta}_{\ell_1+1;1}$ by at most $\frac{K_2}{L^3}$. Because of the partitioning property (3.26), the B-splines with such potentially new coefficients that satisfy the desired inequalities will approximate functions $m|_{[\underline{x},s^0]}(\cdot)$ and $m|_{[s^0,\overline{x}]}(\cdot)$ at the same rate $O\left(\frac{1}{L^3}\right)$ as before.

Now let us show that imposing restrictions $\tilde{\beta}_{L-1;1} = \tilde{\beta}_{2;2} = \tilde{\beta}_{L;1} = \tilde{\beta}_{1;2}$, $\tilde{\beta}_{L_j-2;j} = \tilde{\beta}_{3;j+1}$ that ensure suitable smoothness of the approximation as well as the zero derivative at s^0, does not change the approximation rate.

Indeed, using the approximation properties of the B-splines as well as their derivatives, we have the following sets of properties:

$$\begin{aligned}
\left|\sum_{\ell_j=1}^{L} \tilde{\beta}_{\ell_j;j} p_{\ell_j,L,[s_{j-1},s_j],q}(s^0) - m(s^0)\right| &= O\left(\frac{1}{L^3}\right), \quad j = 1,2,\\
\left|\sum_{\ell_j=1}^{L} \tilde{\beta}_{\ell_j;j} p'_{\ell_j,L,[s_{j-1},s_j],q}(s^0)\right| &= O\left(\frac{1}{L^2}\right), \quad j = 1,2,\\
\left|\sum_{\ell_j=1}^{L} \tilde{\beta}_{\ell_j,[s_{j-1},s_j],q} p''_{\ell_j,L;j}(s^0) - m''(s^0)\right| &= O\left(\frac{1}{L}\right), \quad j = 1,2,
\end{aligned}$$

where the second property also takes into account that $m'(s^0) = 0$.

Note that

$$\begin{aligned}
\sum_{\ell_1=1}^{L} \tilde{\beta}_{\ell_1;1} p_{\ell_1,L,[\underline{x},s^0],q}(s^0) &= \tilde{\beta}_{L;1}, \\
\sum_{\ell_2=1}^{L} \tilde{\beta}_{\ell_2;2} p_{\ell_2,L,[s^0,\overline{x}],q}(s^0) &= \tilde{\beta}_{L;2}, \\
\sum_{\ell_1=1}^{L} \tilde{\beta}_{\ell_1;1} p'_{\ell_1,L,[\underline{x},s^0],q}(s^0) &= \frac{L\triangle\tilde{\beta}_{L;1}}{K_3}, \\
\sum_{\ell_2=1}^{L} \tilde{\beta}_{\ell_2;2} p'_{\ell_2,L,[s^0,\overline{x}],q}(s^0) &= \frac{L\triangle\tilde{\beta}_{2;2}}{K_4}, \\
\sum_{\ell_1=1}^{L} \tilde{\beta}_{\ell_1;1} p''_{\ell_1,L,[\underline{x},s^0],q}(s^0) &= \frac{L^2(2\triangle\tilde{\beta}_{L;1} - \triangle\tilde{\beta}_{L-1;1})}{K_5}, \\
\sum_{\ell_2=1}^{L} \tilde{\beta}_{\ell_2;2} p''_{\ell_2,L,[s^0,\overline{x}],q}(s^0) &= \frac{L^2(\triangle\tilde{\beta}_{3;2} - 2\triangle\tilde{\beta}_{2;2})}{K_6},
\end{aligned}$$

for some constants $K_3 > 0$*,* $K_4 > 0$*,* $K_5 > 0$*,* $K_6 > 0$*.*

These imply that we may have to change the values of coefficients of $\tilde{\beta}_{\ell_1;1}$*,* $\ell_1 = L-2, L-1, L$*, and* $\tilde{\beta}_{\ell_2;2}$*,* $\ell_2 = 1, 2, 3$*, by at most* $\frac{K_7}{L^3}$ *for some* $K_7 > 0$ *to ensure the desired equality constraints as well as to preserve the monotonicity patterns of the approximation. This means (taking into account Eq. (3.26) once again) that with coefficients possibly changed once again, the approximation rate of B-splines is still* $O\left(\frac{1}{L^3}\right)$*.* □

3.4.1 Estimation methodology

We consider the objective function

$$\widehat{Q}^*(s, \beta_{all}, \gamma) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m_{\mathcal{B}}(x_i; s) - z_i'\gamma)^2$$

and solve the problem

$$\min_{s,\beta_{all},\gamma} \widehat{Q}^*(s, \beta_{all}, \gamma) \tag{3.27}$$

subject to the constraints

$$s_1 < s_2 < \ldots < s_J, \tag{3.28}$$

$$\beta_{all} \in T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}. \tag{3.29}$$

In our transformed test statistics, we normalise by the estimate of the variance σ_u^2 of u_i, defined as follows:

$$\breve{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n} \breve{u}_i^2, \tag{3.30}$$

where $\breve{u}_i$ are unconstrained residuals $\breve{u}_i = y_i - \breve{m}_{\mathcal{B}}\left(x_i; \breve{s}\right) - z_i'\breve{\gamma}$ from solving

$$\min_{\beta_{all},\gamma} \widehat{Q}^*\left(\hat{s}, \beta_{all}, \gamma\right)$$

subject to only suitable smoothness constraints in (3.29), without the shape constraints, and with $\hat{s}$ taken from the constrained estimation.

3.5 Testing methodology

3.5.1 Properties of the estimators

To formally prove that our testing procedure works, we need to establish the properties of our estimators. We start by listing regularity conditions.

Condition C3. *(i)* $\{(x_i, z_i', u_i)'\}_{i=1}^n$ *are i.i.d. random vectors. The support of* x *is normalised to* $[0,1]$ *and its density function* $f_X(x)$ *is bounded away from zero on the whole support.* $E(u_i|x_i, z_i) = 0$, $E(u_i^2|x_i, z_i) = \sigma_u^2 < \infty$, u_i *has finite 4th moments, there exists* $\nu > 0$ *such that* $E(|z_i|^{2+\nu}) < \infty$, *and* $E((z_i - E(z_i|x_i))(z_i - E(z_i|x_i))') \neq 0$.

(ii) $m(x)$ *is* $r \geq 3$ *times continuously differentiable.*[5]

(iii) $\frac{(\min_{j=1,\ldots,J+1} L_j)^4}{n} \to 0$, $\frac{(\min_{j=1,\ldots,J+1} L_j)^{2r}}{n} \to \infty$ *as* $n \to \infty$.

Condition $E\left(\left(z_i - E(z_i|x_i)\right)\left(z_i - E(z_i|x_i)\right)'\right) \neq 0$ in C3(i) ensures that no linear combination of z_i can be perfectly predicted by x_i in the least squares sense (we can think of it as no perfect multicollinearity condition: we cannot perfectly substitute between fitting $m_{\mathcal{B}}(x_i)$ and $\gamma' z_i$; adjusting γ cannot fully correct the overall fit if we chose an incorrect switch point). This assumption is needed for identification and root n consistency of the coefficients on z_i. One implication of this assumption is that z_i cannot include a constant. The homoskedasticity assumption could be weakened in a similar way as in Komarova and Hidalgo (2023). Condition C3(ii) on the smoothness of the estimated function determines the quality of *B-spline* approximation. Condition C3(iii) provides the rates at which the number of knots increases to infinity relative to n. This ensures that the bias term (due to the approximation using *B-splines*) is asymptotically negligible.

[5] We could relax this assumption to $r = 2$ and the second derivative is Hölder-continuous.

Consistency

We first show that under the null (3.9) the constrained estimator defined in (3.27)-(3.29) is consistent. To establish this, we consider the regression function $m(\cdot)$ to be a part of a certain compact set and we supplement (3.29) by additional constraints on coefficients $\beta_{\ell-j,j}$s (even though in practice such additional constraints most of the time will not be necessary). We rely on the consistency theorem in Newey and Powell (2003).

Since $m(\cdot)$ is smooth, it is bounded and has a finite Lipschitz constant. We take a very large pointwise bound $A_1 > 0$ and a very large Lipschitz constant A_2 on all the candidate regression functions under consideration (of course, these bounds should be large enough to be true for the underlying regression mean $E[y|x]$). In other words, we take the intersection

$$\Theta_0 = \mathcal{M}_0 \cap \left\{ m(\cdot) \, : \, \sup_{x \in [\underline{x}, \overline{x}]} |m(x)| \leq A_1, \; \sup_{[\underline{x}, \overline{x}]} |m'(x)| \leq A_2 \right\}. \tag{3.31}$$

Proposition 2. *Suppose that conditions C1-C3 hold, $m \in \Theta_0$ and $L_j \to \infty$, $j = 1, \ldots, J+1$, as $n \to \infty$. Then the estimator $\widehat{m}_{\mathcal{B}}(\cdot; \hat{s})$ obtained by solving (3.27)-(3.29) is consistent in the sense that*

$$\sup_{x \in [\underline{x}, \overline{x}]} |\widehat{m}_{\mathcal{B}}(x; \hat{s}) - m(x)| \xrightarrow{p} 0 \quad \textit{as } n \to \infty.$$

The consistency of $\widehat{m}_{\mathcal{B}}(\cdot; \hat{s})$ guarantees the consistency of the switch points, as established in the proposition below.

Corollary 3.5.1. *Under conditions of Proposition 2, the estimators $\hat{s}_j$ of switch points are consistent for s_j, $j = 1, \ldots, J$.*

We can also derive the rates at which the estimators converge to their limits:

Proposition 3. *Under conditions C1-C3:*

$$\hat{\beta} - \beta_0 = O_p\left(\sqrt{\frac{L}{n}}\right), \quad \hat{\gamma} - \gamma_0 = O_p\left(\sqrt{\frac{1}{n}}\right), \quad \hat{s} - s^0 = O_p\left(\frac{L}{\sqrt{n}}\right).$$

3.5.2 The test statistic

Testing procedure and the justification of the need for a transformation

We use a Lagrange multiplier type test.[6] From Assumption 3.2, we know that the true error terms are uncorrelated with any function of the regressors, i.e. $E(u_i f(x_i)) = 0$ for any function $f(\cdot)$. The idea of the test is to check if a similar property is satisfied by the regression residuals:

$$\hat{u}_i = y_i - \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) - \hat{\gamma} z_i. \tag{3.32}$$

[6] For more motivation behind this testing design see the discussion in Komarova and Hidalgo (2023).

A common choice of the function $f(\cdot)$ used in this type of tests, see e.g. Stute (1997), is $f(x_i) = \mathbb{1}(x_i < x)$ for some $x \in [0,1]$, which results in a test statistic of the form:

$$K(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\hat{u}_i. \tag{3.33}$$

Under the null hypothesis, $K(x)$ should be close to zero. However, finding the limiting distribution of this statistic turns out to be problematic. Consider the following expansion:

$$\begin{aligned} K(x) = &\underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)u_i}_{T_0} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right)}_{T_1} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m_{\mathcal{B}}(x_i;\hat{s}) - \widehat{m}_{\mathcal{B}}(x_i;\hat{s})\right)}_{T_2} + \underbrace{(\gamma - \hat{\gamma})' \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)z_i}_{T_3}. \end{aligned}$$

It can be shown that

- $\sqrt{n}T_0 \xrightarrow{d} \sigma\mathcal{B}(F_X(x))$, where $\mathcal{B}(\cdot)$ denotes the standard Brownian motion and $F_X(x)$ is the cdf of x. This term has a well-defined limit which does not depend on the estimates. If this term dominated, we would be able to easily perform tests using standard critical values.

- $T_1 = O_p\left(\frac{1}{\sqrt{n}}\right)$ (follows from Lemma 3.B.6), which implies that this term is not negligible compared to T_0. This is a major difference between our case and that in Komarova and Hidalgo (2023), for whom the term of this form based on the known true s^0 was of a smaller order of magnitude than T_0. We need to modify their transformation to make sure we remove this term as well. An additional complication is that this term is non-linear in parameters, and the Khmaladze transformation relies on linear projections. Because of that, we do not remove this term entirely, but only up to a linear approximation. This is sufficient to ensure that the part which remains after the transformation is of a smaller order and does not affect the limiting distribution.

- $T_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ (follows from Lemma 3.B.5). This is the same rate of convergence as T_0, but unlike T_0 this term does not have a standard known distribution. Instead, the distribution depends on the estimated function in a non-trivial way. The presence of this term motivated the need for a transformation in Komarova and Hidalgo (2023).

- $T_3 = O_p\left(\frac{1}{\sqrt{n}}\right)$ by the standard results on root n convergence of the linear part of a partially linear model, see e.g. Robinson (1988). Hence T_3 has the same convergence rate as the other terms, and its distribution depends on the function we are estimating. We add another modification to the transformation from Komarova and Hidalgo (2023) to remove this term as well.

The last three terms are problematic because their asymptotic distributions depend on the estimated function. As a result, the limiting distribution of the test statistic is not standard. In order to achieve a limiting distribution which would allow us to perform testing using standard techniques, we would like to transform the test statistic in a way which removes the last three terms while leaving the asymptotic behaviour of the first term unchanged. We describe a transformation which achieves this goal in the next section.

The Khmaladze's Transformation

The transformation which removes the problematic terms from $K(x)$ while keeping enough structure of the original statistic to allow for testing is a special case of a martingale transformation introduced by Khmaladze (1982). It can remove all terms linear in some $\widetilde{\boldsymbol{P}}$ we choose, hence we define $\widetilde{\boldsymbol{P}}$ to include: *B-splines* basis functions (these are terms linear in βs, or in other words derivatives with respect to βs: $\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial \beta_l}$, these will remove T_2), zs (derivatives of the linear part with respect to γ_k, these will remove T_3) and linear approximation with respect to s: $\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}$ (this will remove the leading linear component in T_1). All of these are functions of the regressors (x, z), and we assume $E(u_i|x_i, z_i) = 0$, so the residual from regressing u_i on functions of z_i should be very close to u_i, hence the limiting behaviour of the first term should be the same as without a transformation.

In Lemma 3.B.2 we show that:

$$\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s_k} = \mathbb{1}\left[\hat{s}_{k-1}, \hat{s}_k\right) \frac{\hat{s}_{k-1} - x_i}{\hat{s}_k - \hat{s}_{k-1}} \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial x} + \mathbb{1}\left[\hat{s}_k, \hat{s}_{k+1}\right) \frac{x_i - \hat{s}_{k+1}}{\hat{s}_{k+1} - \hat{s}_k} \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial x}$$

where the derivative of *B-spline* is:

$$\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial x} = \sum_{\ell_j=1}^{L} \hat{\beta}_{\ell_j} p'_{\ell_j,L,[\hat{s}_{j-1},\hat{s}_j],q}(x)$$

We let $\mathbf{P}_{L_j;j}(x)$ denote the L_j-dimensional vector of base *B-splines* on $[\hat{s}_{j-1}, \hat{s}_j]$, $j = 1, \ldots, J+1$, computed at x. The estimation uses the long system $\mathbf{P} \equiv (\mathbf{P}_{L_1;1}(x)', \ldots, \mathbf{P}_{L_{J+1};J+1}(x)')'$. However, the constrained estimation under H_0^B results in some binding constraints. Once the binding constraints are enforced, we end up with a smaller system of relevant base *B-splines*.[7]

[7] e.g. we can enforce an equality constraint of the form $\beta_k = \beta_{k+1}$ by replacing the two *B-spline* basis functions p_k and p_{k+1} with a single term of the form $p_k + p_{k+1}$.

We can refer to it as the system of "effective polynomials"and denote it as $\widetilde{\mathbf{P}}(x)$. Let

$$\widetilde{\boldsymbol{P}}_k \equiv \left(\widetilde{\mathbf{P}}(x_k)', z_k', \frac{\partial \widehat{m}_{\mathcal{B}}(x_k;\hat{s})}{\partial s}\right)'.$$

Note that the elements of $\widetilde{\boldsymbol{P}}$ are defined based on the estimates $\hat{\beta}, \hat{s}$ estimated using the entire sample, under the constraints of H_0^B.

In our setting, a transformation $\mathcal{T}$ of a function $W(x)$ can be defined as:

$$(\mathcal{T}W)(x) = W(x) - \int_0^x \widetilde{\mathbf{P}}'(y)\left(\int_x^1 \widetilde{\mathbf{P}}(v)\widetilde{\mathbf{P}}'(v)f_X(v)dv\right)^+ \left(\int_y^1 \widetilde{\mathbf{P}}(w)W(dw)\right) f_X(y)dy \quad (3.34)$$

where A^+ denotes the Moore-Penrose generalised inverse of A. In practice, we cannot evaluate this transformation and instead we use its sample equivalent, $\mathcal{T}_n$. For technical reasons, we add a trimming which removes observations that fall just below knots. Let $\frac{1}{2} < \zeta < 1$ and

$$\mathcal{G} \equiv \left\{ i : x_i \in [0,1] \setminus \bigcup_{\ell}^{L} \left(t_\ell - n^{-\zeta}, t_\ell\right] \right\}. \quad (3.35)$$

where $\{t_\ell\}_{\ell=1}^L$ is the set of knots we use to define our constrained *B-spline* basis functions. We are now ready to define the transformation:

$$(\mathcal{T}_n W)(x) = W(x) - \frac{1}{n}\sum_{i\in\mathcal{G}} \widetilde{\boldsymbol{P}}_i' \left(\frac{1}{n}\sum_{k=1}^n \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}_k' \mathbb{1}(x_k \geq x_i)\right)^+ \int_{x_i}^1 \widetilde{\mathbf{P}}(w)W(dw)\mathbb{1}(x_i < x). \quad (3.36)$$

How the transformation removes the problematic terms

Suppose we apply the transformation to a step function $W(x)$ of the following form:

$$W(x) = \frac{1}{n}\sum_{i=1}^n g(x_i, z_i)\mathbb{1}(x_i < x)$$

where $g(x_i, z_i)$ is some known function. By the properties of a Riemann-Stieltjes integrals with a step function as the integrator:

$$\int_{x_i}^1 \widetilde{\mathbf{P}}(w)W(dw) = \sum_{k=1}^n \widetilde{\boldsymbol{P}}_k \left(\frac{1}{n}g(x_k, z_k)\right)\mathbb{1}(x_k \geq x_i) = \frac{1}{n}\sum_{k=1}^n \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\mathbb{1}(x_k \geq x_i).$$

Then:

$$\begin{aligned}
&(\mathcal{T}_n W)(x) = \\
&= W(x) - \frac{1}{n}\sum_{i\in\mathcal{G}} \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k \mathbb{1}(x_k \geq x_i)\right)^{+} \left(\frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\mathbb{1}(x_k \geq x_i)\right)\mathbb{1}(x_i < x) \\
&= \frac{1}{n}\sum_{i\in\mathcal{G}} \left(g(x_i, z_i) - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k \mathbb{1}(x_k \geq x_i)\right)^{+} \frac{1}{n}\sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\mathbb{1}(x_k \geq x_i)\right)\mathbb{1}(x_i < x) \\
&\quad + \frac{1}{n}\sum_{i\notin\mathcal{G}} g(x_i, z_i)\mathbb{1}(x_i < x) \\
&= \frac{1}{n}\sum_{i\in\mathcal{G}:x_i<x} \left(g(x_i, z_i) - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k\right)^{+} \frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k g(x_k, z_k)\right) \\
&\quad + \frac{1}{n}\sum_{i\notin\mathcal{G}:x_i<x} g(x_i, z_i)
\end{aligned}$$

The term in the first summation is a residual from regressing $g(x_i, z_i)$ on $\widetilde{\boldsymbol{P}}_i$, where the estimator is evaluated using only observations above x_i (i.e. x_k such that $x_k \geq x_i$). The transformed $\mathcal{T}_n W$ at a point x has a similar form to the original W, i.e. it is a weighted sum of functions of the observations x_i below x, but for the majority of indices which fall in $\mathcal{G}$ we use the part of $g(x_i, z_i)$ which cannot be explained by *B-splines* and zs for observations above x_i instead of the whole $g(x_i, z_i)$.

Consider the case where $g(x_i, z_i) = \widetilde{\boldsymbol{P}}'_i a$ for some constant vector a, i.e. where $g(x_i, z_i)$ is a linear combination the constrained *B-spline* functions evaluated at x_i, of z_i and of derivatives of the constrained *B-spline* with respect to the switch point. In this case

$$W(x) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{P}}'_i a\mathbb{1}(x_i < x).$$

Then the transformed version of W is:

$$\begin{aligned}
(\mathcal{T}_n W)(x) &= \frac{1}{n}\sum_{i\in\mathcal{G}:x_i<x} \left(\widetilde{\boldsymbol{P}}'_i a - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k\right)^{+} \frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k a\right) + \frac{1}{n}\sum_{i\notin\mathcal{G}:x_i<x} \widetilde{\boldsymbol{P}}'_i a \\
&= \frac{1}{n}\sum_{i\in\mathcal{G}:x_i<x} \left(\widetilde{\boldsymbol{P}}'_i - \widetilde{\boldsymbol{P}}'_i \left(\frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k\right)^{+} \frac{1}{n}\sum_{k:x_k\geq x_i} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}'_k\right) a + \frac{1}{n}\underbrace{\sum_{i\notin\mathcal{G}:x_i<x} \underbrace{\widetilde{\boldsymbol{P}}'_i a}_{\leq C}}_{=O_p(n^{1-\zeta})} \\
&= 0 + O_p\left(n^{-\zeta}\right) = o_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

The term inside the bracket in the second line is the residual from regressing $\widetilde{\boldsymbol{P}}_i$ on itself, and that residual is identically equal to zero[8] for every i.

This shows that the transformation removes all terms that are linear combinations of constrained *B-splines*, zs and terms linearised in the switch point for $i \in \mathcal{G}$, and as the sample size increases, the number of $i \notin \mathcal{G}$ becomes insignificant. This proves the following results:

Proposition 4. *Let* $T_{2,3}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m_{\mathcal{B}}(x_i;\hat{s}) - \widehat{m}_{\mathcal{B}}(x_i;\hat{s}) + (\gamma - \hat{\gamma})' z_i\right)$. *Then*

$$(\mathcal{T}_n T_{2,3})(x) = o_p\left(n^{-\frac{1}{2}}\right).$$

The distribution of the test statistic

We have shown that the transformation removes the last two terms. The next two results show that the second term becomes negligible and the first term's distribution remains unchanged.

Proposition 5. *Let* $T_4(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}(\hat{s} - s^0)\right)$. *Then*

$$(\mathcal{T}_n T_4)(x) = o_p\left(n^{-\frac{1}{2}}\right)$$

and

$$(\mathcal{T}_n T_1)(x) = o_p\left(n^{-\frac{1}{2}}\right).$$

Proposition 6. *Under C1-C3, the transformation does not affect the limit of* $\sqrt{n}T_0(x)$*:*

$$\sqrt{n}(\mathcal{T}_n T_0)(x) \xrightarrow{weakly} \sigma\mathcal{B}(F_X(x))$$

for any $x \in [0,1]$.

Combining all of these results, we arrive at the pivotal asymptotic distribution of the transformed test statistic.

Theorem 3.5.1. *Under* H_0 *and conditions C1-C3:*

$$\sqrt{n}\left(\mathcal{T}K(x)\right) \xrightarrow{weakly} \sigma_u\mathcal{B}(F_X(x)),$$

$$\breve{\sigma}_u^2 \xrightarrow{p} \sigma_u^2.$$

In order to implement tests based on this asymptotic distribution, we rely on functionals such as Kolmogorov-Smirnov, Cramér-von-Mises and Anderson-Darling, as described in Sec-

[8] For any generalised inverse we can define $P_X = X(X'X)^{-} + X'$, which is a matrix projecting onto the span of X. It has the property that $P_X X = X$, i.e. the projection of X onto X is X. For a given i we let X be the matrix containing columns $\widetilde{\boldsymbol{P}}_i, \widetilde{\boldsymbol{P}}_{i+1}, \ldots, \widetilde{\boldsymbol{P}}_n$ and think of the residual vector from a projection (regression) of X onto itself. The residuals from this regression are zero: $X - P_X X = 0$. The term in the bracket is just the first entry in the residual vector.

tion 3.5.3. The statistics achieve their respective distributions by Theorem 3.5.1 and continuous mapping theorem.

3.5.3 Algorithm outline

STEP 1 Order the sample $\{(x_i, z_i, y_i)\}_{i=1}^n$ in the ascending order of x. Without a loss of generality, we will assume that the original sample is already ordered in this way.

STEP 2 Find a constrained estimator $\widehat{m}_{\mathcal{B}}(\cdot, \hat{s})$ under H_0^B in (3.22) together with estimator $\widehat{\gamma}$ of γ_0 and compute the residuals $\widehat{u}_i = y_i - \widehat{m}_{\mathcal{B}}\left(x_i; \hat{s}\right) - z_i'\widehat{\gamma}$, $i = 1, ..., n$.

Let

$$\widetilde{\boldsymbol{P}}_k \equiv \left(\widetilde{\mathbf{P}}(x_k)', z_k', \frac{\partial \widehat{m}_{\mathcal{B}}(x_k; \hat{s})}{\partial s}\right)'.$$

where $\widetilde{\mathbf{P}}(x)$ denotes the system of "effective polynomials" obtained after enforcing the binding constraints.

STEP 3 For each $i = 1, \ldots, n$, compute the new residual

$$\widehat{v}_i = \widehat{u}_i - \widetilde{\boldsymbol{P}}_i' \left(\sum_{k=1}^n \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}_k' \mathbb{1}(x_k \geq x_i)\right)^+ \sum_{k=1}^n \widetilde{\boldsymbol{P}}_k \mathbb{1}(x_k \geq x_i) \widehat{u}_k. \tag{3.37}$$

STEP 4 Compute the estimate of the variance of u_i, σ_u^2, as $\breve{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^n \breve{u}_i^2$, where $\breve{u}_i$ are unconstrained residuals $\breve{u}_i = y_i - \breve{m}_{\mathcal{B}}\left(x_i; \breve{s}\right) - z_i'\breve{\gamma}$.

STEP 5 Compute $\widetilde{M}_{\tilde{n}}(x_i) = \frac{1}{\sqrt{\tilde{n}}}\sum_{k=1}^{\tilde{n}} \hat{v}_k \mathbb{1}(x_k \geq x_i)$ and calculate the values of standard functionals such as the Kolmogorov-Smirnov, Cramér-von-Mises and Anderson-Darling[9] defined respectively as

$$\mathcal{KS}_{\tilde{n}} = \sup_{i=1,..,n}\left|\frac{\widetilde{M}_n(x_i)}{\breve{\sigma}_u}\right|, \quad \mathcal{CvM}_{\tilde{n}} = \sum_{i=1}^{\tilde{n}} \frac{\widetilde{M}_n(x_i)^2}{n\breve{\sigma}_u^2}, \quad \mathcal{AD}_{\tilde{n}} = \sum_{i=1}^{n} \frac{\widetilde{M}_n(x_i)^2/n}{\breve{\sigma}_u^2 \widehat{F}_X(x_i)}, \tag{3.38}$$

where $\widehat{F}_X$ denotes the empirical c.d.f. of X. Compare them to the critical values $\mathcal{KS}_{\tilde{n}}^*(\alpha_0)$, $\mathcal{CvM}_{\tilde{n}}^*(\alpha_0)$, $\mathcal{AD}_{\tilde{n}}^*(\alpha_0)$, respectively, for a chosen significance level α_0. If. e.g., $\mathcal{KS}_{\tilde{n}} > \mathcal{KS}_{\tilde{n}}^*(\alpha_0)$, reject the null by Kolmogorov-Smirnov at the significance level α_0.

Conducting STEP 1 In the first step we estimate the regression function m under the null hypothesis (3.22). For that we approximate m on each subinterval $[s_{j-1}, s_j]$ by *B-spline* $m_{\mathcal{B};j}$ as defined in (3.13) and the approximation on the whole domain is described by a join

[9] One could, of course, center the process $\widetilde{M}_{\tilde{n}}(x)$ to ensure that it converges to a Brownian bridge indexed by the empirical c.d.f. of X. Then $\mathcal{AD}_{\tilde{n}}$ would be defined in a standard manner as follows: $\mathcal{AD}_{\tilde{n}} = \sum_{i=1}^{\tilde{n}} \frac{\widetilde{M}_{\tilde{n}}(x_i)^2/\tilde{n}}{\breve{\sigma}_u^2 \widehat{F}_X(x_i)(1-\widehat{F}_X(x_i))}$.

in (3.14). The constraints in $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ incorporate (3.17)-(3.18) which guarantee sufficient smoothness.

The fundamental difference between our approach and the approach in Komarova and Hidalgo (2023) is that here we do not take switch points s_j, $j = 1, \ldots, J+1$, as known but estimate them as well.

Due to the requirement (3.7) on the functional properties in classes $\mathcal{M}_j$, $j = 1, \ldots, J+1$, in the overwhelming majority of applications, the properties in each $\mathcal{M}_j$ will be described by conditions on the derivatives of m (potentially on combinations of several derivatives). In cases when each $\mathcal{M}_j$ is described by inequalities on linear combinations of derivatives, all the constraints in $T_{\{(q_j,L_j)\}_{j=1}^{J+1},s}$ are linear inequalities.[10] This was illustrated earlier in the context of the U-shape property in Example 1. Thus, constraints (3.29) in such scenarios are especially easy to implement. However, the optimisation is complicated by the fact that the switch points are unknown. The locations of switch points determine knots points on each subinterval and the values of the polynomials on the *B-spline* bases.

We can see two main approaches to such optimisation. The first approach would be to use the closed-form expressions for *B-spline* base polynomials when programming the objective function in (3.27). These closed form expressions would explicitly account for the knots points which, in their turn, depend on the choice of switch points. Then non-linear optimisation tools can be used.

Another approach, which may especially be convenient when dealing with a small number J of switch points, would be to conduct the grid search. Choose a grid on $[\underline{x}, \overline{x}]$, say of R points, and select all possible J-dimensional subsets from these R points. In these selected subsets J points are naturally ordered and can be treated as candidates for the set of switch points. Then the approximation (3.14) is constructed taking these points as candidate points for partitioning and then the problem (3.27) is solved subject to (3.29) only. In the end we select the sequence of switch point that delivers the smallest value of the objective function. Of course, such a grid search would result in a program conducting the estimation for $\binom{R}{J}$ subsets but, again, may be feasible for small values of J and especially in situations when there is only one switch point.

Conducting STEP 2 In this step we need to define the system of "effective polynomials" by enforcing the biding constraints. Once again, this may be convenient to illustrate using U-shape as an main example. In this case if in the constrained estimate $\widehat{b}_{all}$ of β_{all} we have

$$\widehat{b}_{all,h_1} = \widehat{b}_{all,h_1+1} = \ldots = \widehat{b}_{all,h_2}$$

[10] Equalities, of course, can be represented through inequalities.

for some indices $h_1 < h_2$, and $\widehat{b}_{all,h_1-1} \neq \widehat{b}_{all,h_1}$, $\widehat{b}_{all,h_2} \neq \widehat{b}_{all,h_2+1}$, then instead of $h_2 - h_1 + 1$ different respective base *B-splines* we will include the sum of all these $h_2 - h_1 + 1$ base *B-splines* as one polynomial into $\widetilde{\mathbf{P}}(x)$.

Conducting STEP 3 is straightforward. It comes down to computing the projection of $\{v_k\}_{k=i}^n$ on $\{\widetilde{\mathbf{P}}(x_k)\}_{k=i}^n$ and then using the projection coefficient to compute the new residual for i. This can be conducted by recursive least squares.

Conducting STEP 4 involves finding an unconstrained estimator $\breve{m}_{\mathcal{B}}(x_i)$. This estimator can be found e.g. by either solving

$$\min_{\beta_{all},\gamma} \widehat{Q}^* \left(\hat{s}, \beta_{all}, \gamma\right)$$

subject to only suitable smoothness constraints in (3.29) and with $\hat{s}$ taken from the constrained estimation. Alternatively, one can use just one system of base *B-splines* on the whole interval $[\underline{x}, \overline{x}]$ and conduct unconstrained nonparametric estimation using that base.

Conducting STEP 5 is straightforward.

3.5.4 Bootstrap

Although our test statistic has a pivotal distribution and allows asymptotic testing, the performance may not be the best in small samples. As an alternative, we provide a valid bootstrap algorithm.

STEP 1 Let $\breve{m}_{\mathcal{B}}(x_i, \hat{s})$ and $\breve{\gamma}$ be the estimators analogous to $\widehat{m}_{\mathcal{B}}(x_i; \hat{s})$ and $\hat{\gamma}$ but evaluated without the constraints.[11] Compute the unconstrained residuals as:

$$\breve{\varepsilon}_i = y_i - \breve{m}_{\mathcal{B}}(x_i, \hat{s}) - \breve{\gamma}' z_i.$$

STEP 2 Draw a random sample from the empirical distribution of the unconstrained residuals centred at zero: $\left\{\breve{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^n \breve{\varepsilon}_j\right\}_{i=1}^n$, denote it by $\{\varepsilon_i^*\}_{i=1}^n$. Construct the bootstrap outcomes y_i^* using the constrained estimators:

$$y_i^* = \widehat{m}_{\mathcal{B}}(x_i; \hat{s}) + \hat{\gamma}' z_i + \varepsilon_i^*.$$

STEP 3 Compute the bootstrap estimators $\widehat{m}_{\mathcal{B}}^*(x_i, \hat{s}^*)$ from (3.5.4). Use them to construct

[11] Note that we use the same set of knots as in the constrained case.

the bootstrap residuals

$$\hat{\varepsilon}_i^* = y_i^* - \widehat{m}_{\mathcal{B}}^*(x_i, \hat{s}^*) - \hat{\gamma}^{*'} z_i.$$

Use them to find the value of the bootstrap statistic:

$$\sqrt{n}\,(\mathcal{T}K^*(x))$$
$$= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{G}} \left(\hat{\varepsilon}_i^* - \widetilde{\boldsymbol{P}}_i' \left(\frac{1}{n} \sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \widetilde{\boldsymbol{P}}_k' \mathbb{1}(x_k \geq \widetilde{x}_i) \right)^{+} \frac{1}{n} \sum_{k=1}^{n} \widetilde{\boldsymbol{P}}_k \hat{\varepsilon}_k^* \mathbb{1}(x_k \geq \widetilde{x}_i) \right) \mathbb{1}(x_i < x).$$

Theorem 3.5.2. *Under conditions C1-C3:*

$$\sqrt{n}\,(\mathcal{T}K^*(x)) \xrightarrow{weakly} \sigma_u \mathcal{B}(F_X(x))$$

in probability.

3.6 Monte Carlo simulations

In Scenarios 1-3 we consider

$$y = m(x) + \gamma_0' z + u, \quad u \sim \mathcal{N}(0, \sigma^2)$$

Subscenarios labelled A. We will have no additional covariates – thus, we will take it as given that $\gamma_0 = 0$. We will take x to be uniformly distributed on $[0, 1]$.

Subscenarios labelled B. We will take $\gamma_0 = -2$ and treat γ_0 as unknown in our estimation. We will take x and z to be uniformly distributed on $[0, 1]$ and independent.

Subscenarios labelled C. We will take $\gamma_0 = -2$ and treat γ_0 as unknown in our estimation. We will take x and z to be

$$x = 0.8w_1 + 0.2v,$$
$$z = 0.25 - 0.25w_2 + 0.75v,$$

where w_1, w_2 and v are uniformly distributed on $[0, 1]$ and mutually independent. Thus, on this subscenarios z will be correlated with the base *B-splines.*

Scenario 1. We consider several sub-scenarios within this scenario. Sub-scenarios 1-j,

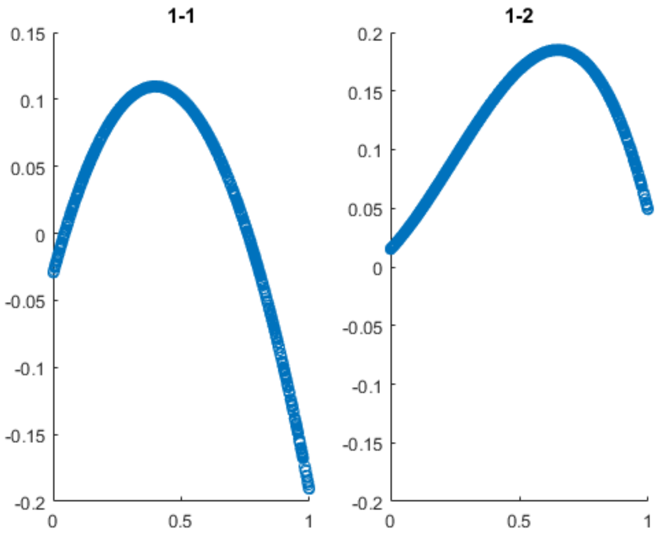
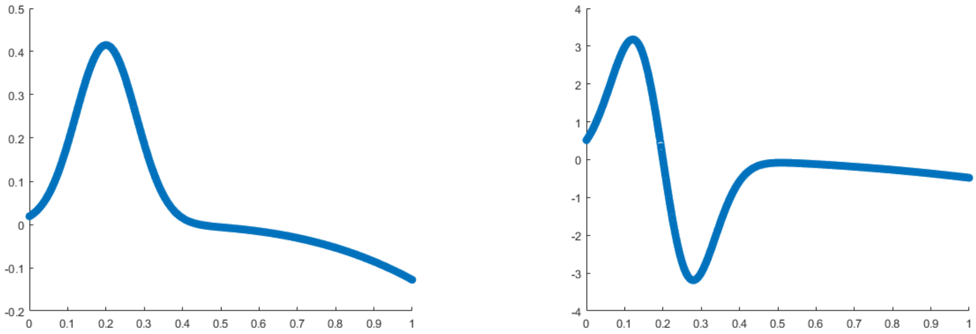


Figure 3.3: Graphs of functions $m(\cdot)$ in Scenarios 1-1 and 1-2.



(a) Function in 1-3.

(b) Derivative of function in 1-3.

Figure 3.4: Graphs of function $m(\cdot)$ and it derivative in Scenario 1-3.

$j = 1, 2, 3$, can be described as

$$m(x) = -0.75(0.2 - x)^2 + 0.415\log(1 + x) \qquad (1\text{-}1),$$

$$m(x) = -0.75(x - 0.5)(0.2 - x)^2 + 0.415\log(1 + x) \qquad (1\text{-}2),$$

$$m(x) = 0.25(0.2 - x)^3 + 0.415\exp(-80(x - 0.2)^2) \qquad (1\text{-}3),$$

and σ is taken to be 0.05 (the findings under H_0 are quite robust with respect to the value of σ).

The graphs of the functions in Scenarios 1-1 and 1-2 are given in Figure 3.3. The graphs of the function in Scenario 1-3 as well as its derivative are given in Figure 3.4.

We start with sub-scenarios 1-jA, $j = 1, 2, 3$, we have $\gamma_0 = 0$ (in other words, there is no control for other covariates).

		A				*B*				*C*			
Setting	Method	B-splines		P-splines		B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
$L'_1 = L'_2 = 4$ h	KS	0.076	0.048	0.088	0.042	0.072	0.028	0.08	0.058	0.0675	0.0375	0.0724	0.05
$N = 1000$	CvM	0.08	0.05	0.092	0.06	0.07	0.02	0.098	0.042	0.0875	0.045	0.09	0.045
$\sigma = 0.05$	AD	0.07	0.042	0.09	0.058	0.086	0.026	0.096	0.038	0.0775	0.04	0.1	0.05
$L'_1 = L'_2 = 6$	KS	0.074	0.03	0.085	0.0325	0.074	0.028	0.1	0.045	0.0575	0.035	0.0825	0.0325
$N = 1000$	CvM	0.074	0.03	0.085	0.035	0.078	0.028	0.0975	0.0525	0.095	0.0475	0.0975	0.05
$\sigma = 0.05$	AD	0.07	0.038	0.0775	0.035	0.07	0.032	0.1	0.0525	0.08	0.04	0.0725	0.045
$L'_1 = L'_2 = 9$	KS	0.088	0.052	0.085	0.045	0.068	0.03	0.098	0.058	0.0925	0.045	0.0925	0.045
$N = 1000$	CvM	0.102	0.048	0.095	0.05	0.072	0.038	0.098	0.06	0.095	0.065	0.095	0.065
$\sigma = 0.05$	AD	0.088	0.048	0.0875	0.045	0.076	0.034	0.09	0.056	0.0925	0.0575	0.0925	0.575

Table 3.1: Test for an inverse U-shape in Scenario 1-1.

		A				*B*				*C*			
Setting	Method	B-splines		P-splines		B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
$L'_1 = L'_2 = 4$	KS	0.052	0.018	0.064	0.024	0.074	0.026	0.095	0.0525	0.0625	0.025	0.08	0.0275
$N = 1000$	CvM	0.066	0.028	0.084	0.042	0.086	0.038	0.095	0.0425	0.0625	0.035	0.0875	0.0375
$\sigma = 0.05$	AD	0.068	0.032	0.094	0.032	0.084	0.03	0.08	0.0525	0.06	0.0225	0.095	0.0425
$L'_1 = L'_2 = 6$	KS	0.074	0.028	0.0925	0.0425	0.066	0.024	0.09	0.048	0.0825	0.05	0.1075	0.05
$N = 1000$	CvM	0.078	0.03	0.0925	0.0525	0.056	0.026	0.072	0.034	0.1175	0.0475	0.1	0.0475
$\sigma = 0.05$	AD	0.072	0.028	0.085	0.0425	0.054	0.02	0.074	0.046	0.105	0.0525	0.075	0.0375
$L'_1 = L'_2 = 9$	KS	0.09	0.046	0.09	0.046	0.09	0.05	0.085	0.0475	0.115	0.07	0.1	0.0475
$N = 1000$	CvM	0.088	0.052	0.088	0.052	0.106	0.048	0.0875	0.045	0.1225	0.0725	0.0925	0.0475
$\sigma = 0.05$	AD	0.09	0.05	0.09	0.05	0.086	0.062	0.0925	0.05	0.1325	0.08	0.0975	0.0425

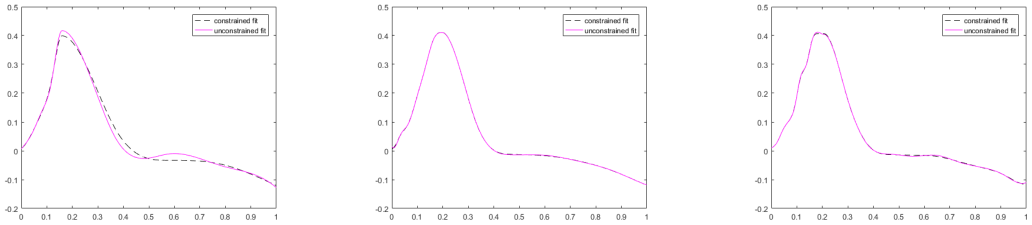
Table 3.2: Test for an inverse U-shape in Scenario 1-2.

We apply our *B-spline* and *P-spline* methodology to test an inverse U-shape in m. Results are given in Tables 3.1-3.3.

A particularly interesting case in this setting is the testing result in Table 3.3 where we see drastically different results for $L'_1 = L'_2 = 4$ compared to other cases of $L'_1 = L'_2 = 6$ and $L'_1 = L'_2 = 9$. The intuition for this can be obtained from Figure 3.4, where we see that the derivative of function m is close to constant on a subinterval. Since our bootstrap draws residuals from the *unconstrained B-splines* fit, the drastic differences between unconstrained and constrained fits in that subinterval can create the high rejection rate. The typical *B-splines* fits with an adaptive choice of the turning point for our three cases of (L'_1, L'_2) are given in Figure 3.5. What we see is that for the case $L_1 = L'_2 = 4$ the unconstrained *B-splines* typically

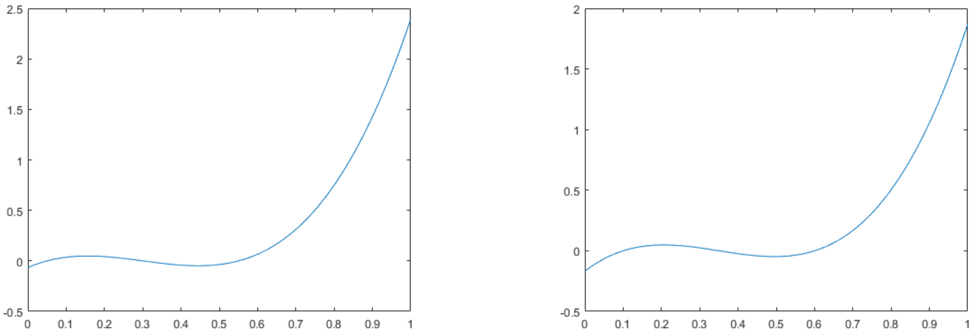
		A				*B*				*C*			
Setting	Method	B-splines		P-splines		B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
$L'_1 = L'_2 = 4$	KS	1	1	0.745	0.4675	1	0.998	0.72	0.5025	1	1	0.33	0.0675
$N = 1000$	CvM	1	1	0.6075	0.4275	1	0.998	0.5425	0.365	1	1	0.295	0.0575
$\sigma = 0.05$	AD	1	0.996	0.62	0.4375	1	1	0.5	0.2925	1	0.995	0.1025	0.0375
$L'_1 = L'_2 = 6$	KS	0.082	0.028	0.1	0.044	0.086	0.042	0.095	0.06	0.1025	0.05	0.0975	0.045
$N = 1000$	CvM	0.084	0.04	0.098	0.048	0.078	0.038	0.0925	0.045	0.1	0.045	0.1	0.045
$\sigma = 0.05$	AD	0.084	0.044	0.096	0.044	0.076	0.024	0.0975	0.04	0.0775	0.05	0.0825	0.0475
$L'_1 = L'_2 = 9$	KS	0.124	0.052	0.08	0.0375	0.084	0.036	0.0775	0.035	0.085	0.0275	0.0875	0.0425
$N = 1000$	CvM	0.132	0.052	0.095	0.0425	0.092	0.054	0.09	0.0525	0.0625	0.04	0.0875	0.055
$\sigma = 0.05$	AD	0.118	0.058	0.09	0.05	0.092	0.04	0.09	0.0475	0.065	0.025	0.0925	0.04

Table 3.3: Test for an inverse U-shape in Scenario 1-3.



(a) $L'_1 = L'_2 = 4$ (b) (c) $L'_1 = L'_2 = 9$

Figure 3.5: Typical *B-spline* fits of function $m(\cdot)$ in Scenario 1-3.



(a) $a = 0.05$. (b) $a = 0.1$.

Figure 3.6: Graphs of functions in Scenario 2.

estimates the function as being increasing on a part of the subinterval with the derivative close to zero, which explains rejection rates for that case in Table 3.3. This situation is no longer the case when $L'_1 = L'_2 = 6$ or $L'_1 = L'_2 = 9$, as can be seen from typical fits in 3.5 as well.

Scenario 2. In Scenario 2 we use $m(x) = x - a - 6(x-a)^2 + 8(x-a)^3$. The graphs of this function for $a = 0.05$ and $a = 0.1$ are given in Figure 3.6. As we can see, the functions are not U-shaped but it is a difficult case to reject U-shape as its violations only happen in a small domain near one of the support ends. It is harder to reject U-shape for $a = 0.05$ than for $a = 0.1$.

3.7 Applications

3.7.1 "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development", by Q.Ashraf and O. Galor, American Economic Review, 2013

In our first application we look at the data from Ashraf and Galor (2013). The paper argues that in the course of the prehistoric exodus of Homo Sapiens out of Africa, genetic diversity has had a persistent hump-shaped effect on the the logarithm of population density and on comparative economic development. The paper contains many findings related to the presence

		$a = 0.05$				$a = 0.1$			
Setting	Method	B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%
$L_1' = 4, L_2' = 4$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.998	0.992	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 6, L_1' = 6$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	1	1	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 9, L_2' = 9$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	1	1	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 4, L_2' = 4$	KS	0.792	0.654	0.8475	0.795	1	1	1	1
$N = 1000$	CvM	0.636	0.448	0.7425	0.64	1	1	1	1
$\sigma = 0.1$	AD	0.908	0.812	0.92	0.88	1	1	1	1
$L_1' = 6, L_1' = 6$	KS	0.858	0.756	0.89	0.785	1	1	1	1
$N = 1000$	CvM	0.786	0.664	0.83	0.755	1	1	1	1
$\sigma = 0.1$	AD	0.95	0.904	0.95	0.925	1	1	1	1
$L_1' = 9, L_2' = 9$	KS	0.874	0.78	0.865	0.8025	1	1	1	1
$N = 1000$	CvM	0.8	0.678	0.8325	0.77	1	0.992	1	1
$\sigma = 0.1$	AD	0.964	0.934	0.9525	0.94	1	1	1	1
$L_1' = 4, L_2' = 4$	KS	0.132	0.066	0.164	0.096	0.748	0.536	0.732	0.55
$N = 1000$	CvM	0.144	0.084	0.186	0.1	0.688	0.486	0.668	0.564
$\sigma = 0.25$	AD	0.25	0.152	0.262	0.158	0.876	0.714	0.872	0.764
$L_1' = 6, L_1' = 6$	KS	0.188	0.086	0.198	0.136	0.794	0.62	0.8075	0.6725
$N = 1000$	CvM	0.196	0.11	0.246	0.16	0.73	0.544	0.73	0.5675
$\sigma = 0.25$	AD	0.286	0.17	0.32	0.228	0.876	0.738	0.8625	0.7675
$L_1' = 9, L_2' = 9$	KS	0.17	0.106	0.208	0.136	0.65	0.526	0.672	0.532
$N = 1000$	CvM	0.2	0.102	0.24	0.166	0.576	0.39	0.548	0.438
$\sigma = 0.25$	AD	0.278	0.202	0.324	0.222	0.728	0.58	0.728	0.596

Table 3.4: Test for U-shape in Scenario 2.

		$a = 0.05$				$a = 0.1$			
Setting	Method	B-splines		P-splines		B-splines		P-splines	
		10%	5%	10%	5%	10%	5%	10%	5%
$L_1' = 4, L_2' = 4$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.996	0.992	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 6, L_1' = 6$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.998	0.998	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 9, L_2' = 9$	KS	1	1	1	1	1	1	1	1
$N = 1000$	CvM	0.99	0.9475	1	1	1	1	1	1
$\sigma = 0.05$	AD	1	1	1	1	1	1	1	1
$L_1' = 4, L_2' = 4$	KS	0.796	0.65	0.904	0.83	1	1	1	1
$N = 1000$	CvM	0.672	0.486	0.864	0.752	1	1	1	1
$\sigma = 0.1$	AD	0.91	0.814	0.954	0.902	1	1	1	1
$L_1' = 6, L_1' = 6$	KS	0.78	0.662	0.8725	0.8075 1	1	1	1	
$N = 1000$	CvM	0.66	0.522	0.8175	0.7625	1	0.998	1	1
$\sigma = 0.1$	AD	0.888	0.822	0.9475	0.9175	1	1	1	1
$L_1' = 9, L_2' = 9$	KS	0.648	0.448	0.71	0.535	1	1	1	1
$N = 1000$	CvM	0.544	0.366	0.6775	0.4825	1	1	1	1
$\sigma = 0.1$	AD	0.816	0.676	0.8325	0.7057	1	1	1	1
$L_1' = 4, L_2' = 4$	KS	0.162	0.09	0.195	0.1125	0.662	0.484	0.7725	0.65
$N = 1000$	CvM	0.184	0.118	0.2075	0.1375	0.662	0.468	0.6825	0.5575
$\sigma = 0.25$	AD	0.276	0.16	0.2575	0.19	0.818	0.684	0.8225	0.75
$L_1' = 6, L_1' = 6$	KS	0.138	0.068	0.195	0.1	0.622	0.406	0.6775	0.4475
$N = 1000$	CvM	0.15	0.096	0.1875	0.105	0.594	0.352	0.6075	0.4675
$\sigma = 0.25$	AD	0.214	0.122	0.245	0.14	0.722	0.47	0.7725	0.6125
$L_1' = 9, L_2' = 9$	KS	0.184	0.086	0.245	0.1725	0.6	0.436	0.6475	0.5175
$N = 1000$	CvM	0.172	0.1	0.2625	0.1725	0.422	0.314	0.5725	0.4775
$\sigma = 0.25$	AD	0.278	0.162	0.31	0.185	0.62	0.47	0.6775	0.5725

Table 3.5: Test for U-shape in Scenario 2-B.

Figure 3.7: Table 4 from Ashraf and Galor (2013).

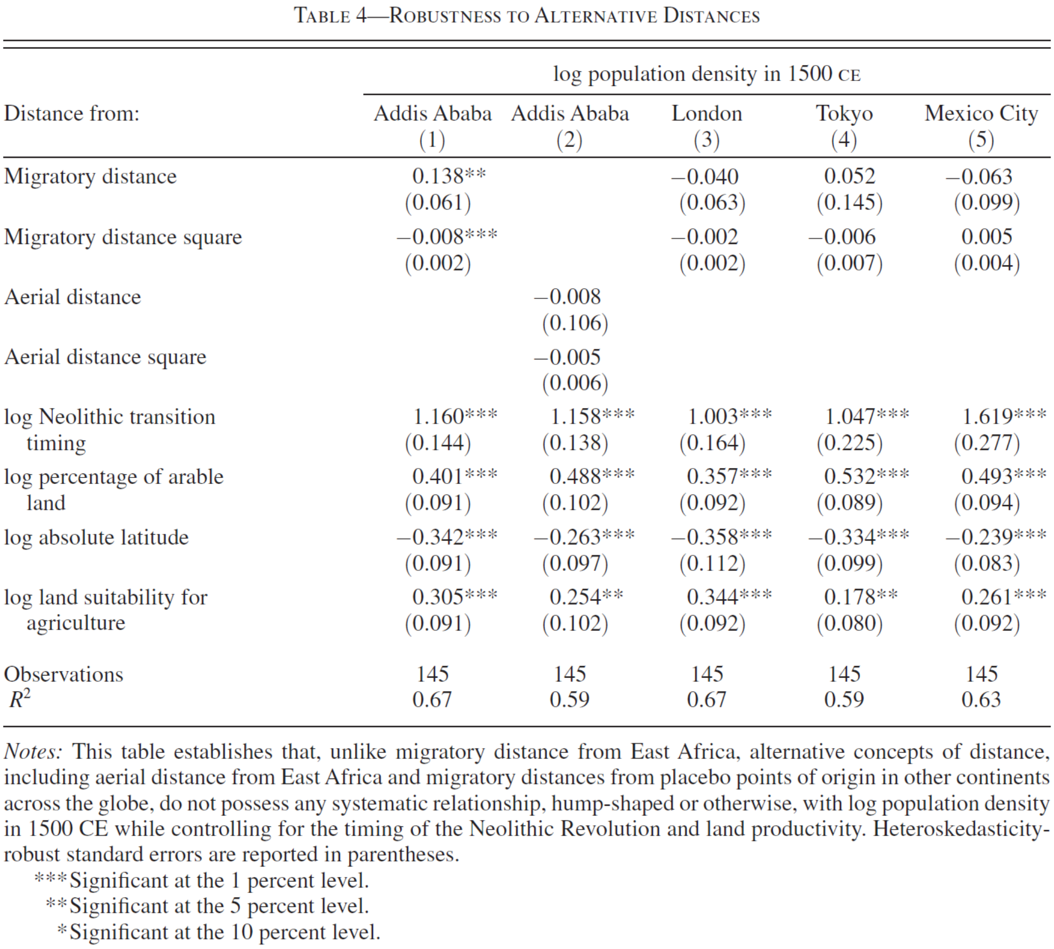


Table 4—Robustness to Alternative Distances

	log population density in 1500 CE				
Distance from:	Addis Ababa (1)	Addis Ababa (2)	London (3)	Tokyo (4)	Mexico City (5)
Migratory distance	0.138** (0.061)		−0.040 (0.063)	0.052 (0.145)	−0.063 (0.099)
Migratory distance square	−0.008*** (0.002)		−0.002 (0.002)	−0.006 (0.007)	0.005 (0.004)
Aerial distance		−0.008 (0.106)			
Aerial distance square		−0.005 (0.006)			
log Neolithic transition timing	1.160*** (0.144)	1.158*** (0.138)	1.003*** (0.164)	1.047*** (0.225)	1.619*** (0.277)
log percentage of arable land	0.401*** (0.091)	0.488*** (0.102)	0.357*** (0.092)	0.532*** (0.089)	0.493*** (0.094)
log absolute latitude	−0.342*** (0.091)	−0.263*** (0.097)	−0.358*** (0.112)	−0.334*** (0.099)	−0.239*** (0.083)
log land suitability for agriculture	0.305*** (0.091)	0.254** (0.102)	0.344*** (0.092)	0.178** (0.080)	0.261*** (0.092)
Observations	145	145	145	145	145
R^2	0.67	0.59	0.67	0.59	0.63

Notes: This table establishes that, unlike migratory distance from East Africa, alternative concepts of distance, including aerial distance from East Africa and migratory distances from placebo points of origin in other continents across the globe, do not possess any systematic relationship, hump-shaped or otherwise, with log population density in 1500 CE while controlling for the timing of the Neolithic Revolution and land productivity. Heteroskedasticity-robust standard errors are reported in parentheses.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

or absence of hump-shaped effects. The authors use quadratics in all their specifications to establish the presence or absence of hump shapes.

We apply our method and compare our results to Ashraf and Galor (2013) Table 4, which contains robustness checks to using alternative distances. It is given in Figure 3.7. The authors' conclusion is that "the results presented in Table 4 indicate that migratory distance from East Africa is the only concept of distance that confers a significant nonmonotonic effect on log population density." We want to analyse and assess these findings using our methodology.

Our first series of tests is about the specification

$$\ln pd1500 = \alpha + \beta dist + \gamma dist^2 + z'\delta + u, \tag{3.39}$$

where *dist* is a distance notion from Figure 3.7, z is the set of 4 controls used there in every column, and u is the error term.

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
Column (1)	<	<	<	<	<	<	<	<	<
Column (2)	<	<	<	>	>	>	>	>	<
Column (3)	>	>	>	>	>	>	>	>	>
Column (4)	<	<	<	>	>	>	>	<	<
Column (5)	>	>	<	>	>	<	<	<	<

Table 3.6: Ashraf and Galor (2013). Test for quadratic specifications in Table 4 in Ashraf and Galor (2013). The percentages (90%, 95%, 99%) stand for the different critical values. All critical values are based on 1000 bootstrap draws. > (<) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

Finding 1, The quadratic specification in (3.39) is rejected at the 5% significance level for Columns 2-5 in Table 4 in Ashraf and Galor (2013).

For this finding we use an approach based on Khmaladze's transformation but within the context of a semiparametric regression (rather than a nonparametric one) as discussed in Stute, Thies, and Zhu (1998). Based on that approach, quadratic specifications in distance (plus other covariates) in Table 4 in Ashraf and Galor (2013) are rejected for Columns (2)-(5) at the 5% level by at least one of our testing functionals. More detailed results are given in Table 3.6, where we can see that for Columns (2), (3) and (5) the quadratic specifications are rejected by at least two functionals we employ (for Column (3) it is rejected by all three functionals). Results for Column (1), thus, can be taken as supportive of Ashraf and Galor (2013) findings for that particular specification, which cannot be said for specifications in other columns used to justify the use of one particular migratory distance in Column (1).

An immediate conclusion here is that robustness to alternative distances needs to be analysed through more general hump-shapes that go beyond quadratics. This naturally brings us to using our method.

For a distance of interest in a respective column we choose cubic *B-splines* on both sides of a candidate switch point with intervals on both sides being uniformly divided into 4 subintervals. This results in 12 base splines overall but the constraints of smoothness of the function at the switch point effectively reduce this number of unknown parameters with respect to the distance variable to 9 (for comparison, in the quadratic specification it is 3 unknown parameters). Table 3.7 shows the results of performing the test using our method with *B-splines*. As we can see, for models analogous to those in Ashraf and Galor (2013) Table 4 which differ from them only in a more general specification with respect to a distance variables, a hump-shape relation with respect to distance is not rejected for distances and in all of the columns.

These conclusions are very different from those reached by quadratic specifications used in Ashraf and Galor (2013). Namely, the aerial distance from East Africa and migratory distance

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
Column (1)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (2)	$>$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (3)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (4)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Column (5)	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$

Table 3.7: Ashraf and Galor (2013) data. *B-splines* based test for hump-shaped specifications in Table 4 in Ashraf and Galor (2013). All critical values are based on 1000 bootstrap draws. $>$ ($<$) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

	difference 1	95% CI	difference 2	95% CI
Column (1)	1.3061	(0.3611,2.8020)	-3.2675	(-4.1464, -2.2730)
Column (2)	0.9621	(0.1208, 2.4071)	-1.2410	(-2.2153,-0.8581)
Column (3)	0.2132	$(1.7 \cdot 10^{-13}, 1.0135)$	-2.9977	(-3.8233,-2.0648)
Column (4)	0.4352	(0.0044,1.8457)	-3.3684	(-4.7242,-2.2614)
Column (5)	1.3980	(1.0919,2.7541)	-0.4366	$(-2.2800, -1.5 \cdot 10^{-13})$

Table 3.8: Ashraf and Galor (2013) data. Analysis whether there are statistically significant changes in the hump-shaped *B-splines* fit before the estimated switch point and also after it. All critical values are based on 1000 bootstrap draws.

from Tokyo have systematic hump-shaped effect on the logarithm of population density in 1500 CE.

A reader may say that our approach to testing hump-shaped relationship potentially allows only weak monotonicity on both sides of the turning points and, thus, potentially hump-shaped relations we find could exhibit a constant effect before or after the estimated turning point. To address this, we look at our *B-splines* hump-shaped fit, compute (a) the difference between the fitted value at the lowest value of the distance and the fitted value at the switch point; (b) the difference between the fitted value at the switch point and the fitted value at the largest value of the distance, and then we construct a 95% bootstrap confidence intervals for both these differences. The results are given in Table 3.8 and allow us to conclude that for Columns (1), (2) and (4) both parts of the fitted curve are strictly monotone at the 5% significance level. For column (3) the first part (increasing) is not rejected to be constant and for Column (5) the second part (decreasing) is not rejected to be constant. Since our constrained estimation imposes difference 1 to be non-negative and difference 2 to be non-positive, what what be more informative for Columns (3) and (5) is the percentage of analogous bootstrap differences that are close to 0. In the case of Column (5), difference 2 is within 10^{-6} distance from 0 in 5.9% samples (so, 90% CI would have 0 on the boundary as well). At the 5% significance level ,the fitted function has a strict increase over the domain before the estimated switch point and a strict decreases after it.

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
Column (1)	<	<	<	<	<	<	<	<	<
Column (2)	<	<	<	<	<	<	<	<	<
Column (3)	>	<	<	<	<	<	<	<	<
Column (4)	<	<	<	<	<	<	<	<	<
Column (5)	<	<	<	<	<	<	<	<	<

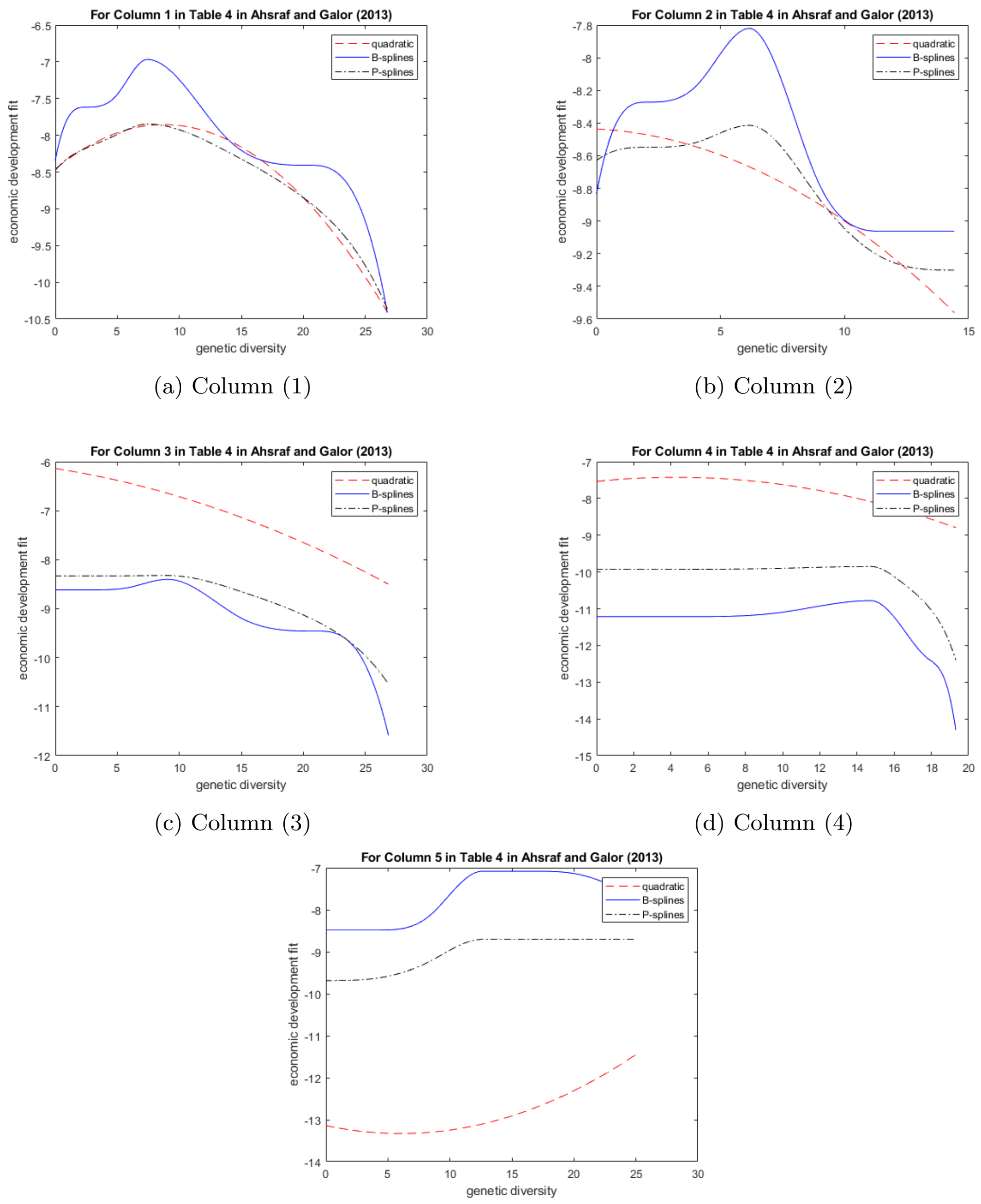
Table 3.9: Ashraf and Galor (2013) data. *P-splines* based test for hump-shaped specifications in Table 4 in Ashraf and Galor (2013). All critical values are based on 1000 bootstrap draws. > (<) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

Finally, taking into account the small size of the sample (just 145 observations) and the presence of additional controls in some specifications in Table 4, we use *P-splines* that is an effective tool for dealing with potential overfitting and avoiding fitted lines that are "too wiggly." For *P-splines*, we penalise second differences of coefficients choosing the same penalty on different sides of the switch point. The penalty is chosen by the cross validation approach in Eilers and Marx (1996). If for a model the penalty is rather large, then the fitted regression mean would have a shape closer to a quadratic one.

Test results using *P-splines* are given in Table 3.9. The substantive conclusions are largely similar to those in Table 3.7.

Finally, we present the following fitted curves for all the columns: first, obtained by quadratic specification in Ashraf and Galor (2013); second, obtained by our *B-spline* methodology under the hump-shape constraint; third, obtained by our *P-splines* methodology with cross-validated penalties enforcing the hump-shape constraint, these are contained in Figure 3.8.

As we can see for the model in Column (1), the fit by *P-splines* is similar to the one provided by the quadratic function. However, for other columns the results are very different. For the model in Column (2), the quadratic specification gives us a monotonically decreasing fit on the domain of the distance, whereas both nonparametric fits indicate a hump-shaped pattern (recall that they are not rejected by either *B-splines* or *P-splines*) with visible asymmetries around the turning point. For the model in Column (4) both non-parametric fits indicate a turning point much further to the right than that given by the quadratic fit. Also, in either non parametric fit the decrease after the turning point is much sharper compared to the increase before that (for *P-splines* the curve before the turning point looks almost flat even though statistically it is not). For the model in Column (5), the quadratic specification fit is U-shaped rather than hump-shaped (recall that in Table 4 in Ashraf and Galor (2013) that it is statistically insignificant at the 5% level) which is drastically different from the hump-shaped nonparametric fits exhibiting visible asymmetry around the turning point.



(a) Column (1)

(b) Column (2)

(c) Column (3)

(d) Column (4)

(e) Column (5)

Figure 3.8: Fitted curves for models in AG Table 4.

In summary, our methodology finds relationship between migratory distance and the log population density in 1500 CE in to be monotonic for specifications in Columns (3) and (5) (at the 5% level). Our findings for Column (1), including the estimation results by *P-splines*, are largely consistent with Ashraf and Galor (2013). Our findings for models and distances in Columns (2)-(5) are different from those in Ashraf and Galor (2013). Namely, in columns (2) and (4) we find hump-shaped relationship between migratory distances and the log population density in 1500 CE and they are different from quadratic ones. In Column (3) we do not reject at 5% level that find a monotonic weakly decreasing relationship, which is consistent with Ashraf and Galor (2013). However, we do have a statistically significant change in the monotonic relationship if we compare the values of our fitted function at the lower and upper support points (this is different from lack of statistical significance conclusions in Ashraf and Galor (2013)). In Column (5) we do not reject at 5% level that find a monotonic weakly increasing relationship and we also find the change in this monotone function over the domain to be statistically significant, with both of these features being different from findings in Ashraf and Galor (2013). These differences are best explained by the fact that in Columns (2)-(5) the best fitted curves under the null of a hump-shape exhibit striking asymmetries around the turning points which is not allowed by quadratic specifications.

3.7.2 Child penalty

We consider the country-level model:[12]

$$\begin{gathered} \mathit{Child\_Penalty}_i = m(\log(\mathit{GDP\_per\_capita}_i)) + \beta \mathit{Employment\_Gap}_i + u_i \qquad (3.40) \\ E[u_i | \log(\mathit{GDP\_per\_capita}_i), \mathit{Employment\_Gap}_i] = 0, \end{gathered}$$

and test

$$H_0 : m \quad \text{is hump-shaped.}$$

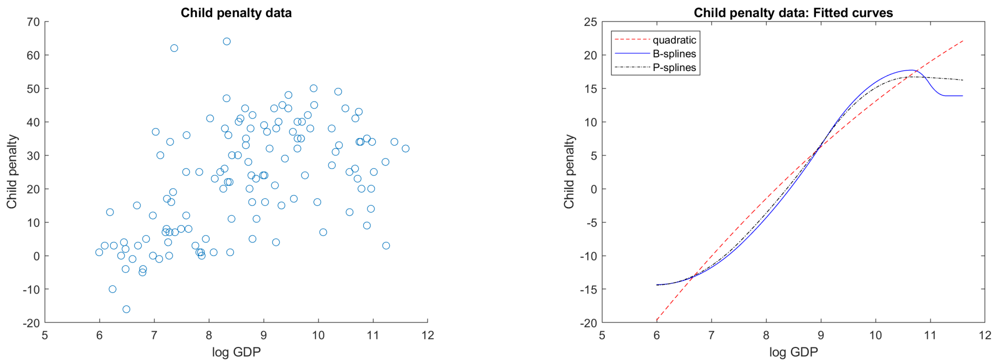
The employment gap between women and men can reflect societal norms, policies, and labour market dynamics that influence the child penalty. Larger employment gaps e.g. might indicate less support for working mothers, which could exacerbate the child penalty.

The left panel of Figure 3.9 plots the data $(\log(\mathit{GDP\_per\_capita}), \mathit{Child\_Penalty})$ and the right hand plots the fitted curves $m(\log(\mathit{GDP\_per\_capita}))$ obtained by (a) a quadratic specification $m(\log(\mathit{GDP\_per\_capita})) = \gamma_0 + \gamma_1 \mathit{Child\_Penalty} + \gamma_2 \mathit{Child\_Penalty}^2$, (b) *B-spline* specification for $m(\cdot)$, and (c) *B-spline* specification for $m(\cdot)$ estimated with the use of penalty on the second-differences of coefficients as explained earlier (so *P-splines*).

As we can see, the quadratic specification finds a strictly increasing curve within the domain of log of GDP per capita. We start by applying our test for testing the quadratic specification of $m(\cdot)$ in (3.40) analogously to how it was conducted in the previous application. The results are in Table 3.10. As we can see, all three types of tests reject a quadratic form of $m(\cdot)$ at the 5% significance level. Therefore, quadratics do not look like a suitable approach in capturing a nonlinear relationship between log of GDP per capita and child penalty.

Our next step is to test the null hypothesis of hump-shape using *B-splines* and *P-splines* approach. We choose quadratic *B-splines* on both sides of a candidate switch point with

[12] We are grateful to Camille Landais and Gabriel Leite-Mariante for providing us with the data.



(a) Data

(b) Fitted curves

Figure 3.9: Left panel: data on log of GDP per capita and child penalty. Right panel: Fitted curves for the model (3.40).

Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
>	>	>	>	>	>	>	>	<

Table 3.10: Child penalty data. Test for a quadratic form of $m(\cdot)$ in (3.40). "cv" stands for the critical value. All critical values are based on 1000 bootstrap draws. > (<) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

intervals on both sides being uniformly divided into 4 subintervals. This results in 10 base splines overall but the constraints of smoothness of the function at the switch point effectively reduce this number of unknown parameters with respect to the distance variable to 7 (compared to three unknown parameters in a quadratic specification). The results are given in 3.11.

As we can see, the null of a hump-shaped relationship is not rejected at the 10% significance level. The switch point is found to be 10.67 (on the grid of equidistant 1001 grid points in the domain of log of GDP per capita).

3.8 Conclusion

This paper develops a robust nonparametric methodology for testing shape constraints in regression analysis, accommodating multiple shape changes across the domain of the regressor.

	Kolmogorov-Smirnov			Cramer-von-Mises			Anderson-Darling		
	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv	90% cv	95% cv	99% cv
B-splines	<	<	<	<	<	<	<	<	<
P-splines	<	<	<	<	<	<	<	<	<

Table 3.11: Child penalty data data. *B-splines* and *P-splines* based tests for hump-shaped $m(\cdot)$ in (3.40). All critical values are based on 1000 bootstrap draws. > (<) means that the test statistic for the functional indicated in the first row is greater (is less) than the respective critical value for that functional.

Our approach extends beyond conventional U-shaped or hump-shaped patterns to a broad class of nonlinear shapes, including S-shapes, W-shapes, etc. Unlike previous methods that rely on parametric assumptions or require predetermined switch points, our approach identifies turning points adaptively within the data. This allows for greater flexibility and more accurate representation of complex nonlinear relationships, which are often misrepresented by simplistic parametric polynomial (in particular, quadratic) models.

The theoretical contributions of this paper include ensuring that the adaptive estimation of turning points does not compromise the statistical properties of the test statistics, both in finite samples and asymptotically. Practically, the methodology improves the power and interpretability of shape testing by reducing reliance on restrictive parametric forms. As our applications demonstrate, standard parametric approximations can miss or distort true underlying relationships, while our method captures these dynamics more precisely.

In summary, this paper provides a valuable tool for researchers across disciplines who require a flexible, rigorous approach to testing complex shape constraints. The methodology broadens the scope of nonparametric analysis in regression contexts, offering a unified framework that can be applied to partially linear models (or partially parametric models more generally) and expanded to incorporate multiple turning points. Future research may build on this work by further refining the estimation of turning points and exploring additional applications in diverse empirical settings.

Bibliography for Chapter 3

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. 2005. "Competition and innovation: An inverted-U relationship." *The quarterly journal of economics* 120 (2): 701–728.

Aghion, Philippe, John Van Reenen, and Luigi Zingales. 2013. "Innovation and institutional ownership." *American economic review* 103 (1): 277–304.

Andrews, Donald. 1997. "A Conditional Kolmogorov Test." *Econometrica* 65 (5): 1097–1128.

Ashraf, Quamrul, and Oded Galor. 2013. "The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development." *American Economic Review* 103 (1): 1–46.

Bessec, Marie, and Julien Fouquau. 2008. "The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach." *Energy Economics* 30 (5): 2705–2721.

Blanchflower, David G. 2020. *Is Happiness U-shaped Everywhere? Age and Subjective Well-being in 132 Countries.* Working Paper, Working Paper Series 26641. National Bureau of Economic Research, January.

Brown, Robert L, James Durbin, and James M Evans. 1975. "Techniques for testing the constancy of regression relationships over time." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 37 (2): 149–163.

Calabrese, E.J., and L.A. Baldwin. 2001. "Hormesis: U-shaped dose responses and their centrality in toxicology."

Chen, Xiaohong. 2007. "Large sample sieve estimation of semi-nonparametric models." *Handbook of econometrics* 6:5549–5632.

Clark, Andrew. 2007. "Born to Be Mild? Cohort Effects Don't (Fully) Explain Why Well-Being is U-Shaped in Age."

Das, Somnath, and Baruch Lev. 1994. "Nonlinearity in the Returns-Earnings Relation: Tests of Alternative Specifications and Explanations." *Contemporary Accounting Research* 11 (1): 353–379.

De Boor, Carl. 1978. *A practical guide to splines.* Vol. 27. springer-verlag New York.

Delgado, Miguel A., and Javier Hidalgo. 2000. "Nonparametric inference on structural breaks." *Journal of Econometrics* 96 (1): 113–144.

Eddy, William F. 1980. "Optimum Kernel Estimators of the Mode." *Annals of Statistics* 8 (4): 870–882.

Eilers, Paul H. C., and Brian D. Marx. 1996. "Flexible smoothing with B-splines and penalties." *Statistical Science* 11 (2): 89–121.

Engle, Robert F., C. W. J. Granger, John Rice, and Andrew Weiss. 1986. "Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Journal of the American Statistical Association* 81 (394): 310–320.

Feder, Paul I. 1975. “On Asymptotic Distribution Theory in Segmented Regression Problems–Identified Case.” *Annals of Statistics* 3 (1): 49–83.

Freeman, Robert N., and Senyo Y. Tse. 1992. “A Nonlinear Model of Security Price Responses to Unexpected Earnings.” *Journal of Accounting Research* 30 (2): 185–209.

Ganz, Scott C. 2024. *A Goldilocks Approach to Testing Nonmonotonic Hypotheses in Management and Strategy.* Working Paper, Working Paper Series. Georgetown McDonough School of Business and AEI.

Goldin, Claudia. 1995. “The U-Shaped Female Labor Force Function in Economic Development and Economic History.” In *Investment in Women’s Human Capital and Economic Development,* edited by T. P. Schultz, 61–90. University of Chicago Press.

Groes, Fane, Philipp Kircher, and Iourii Manovskii. 2014. “The U-Shapes of Occupational Mobility.” *The Review of Economic Studies* 82 (2): 659–692.

Heckman, Nancy E. 1986. “Spline Smoothing in a Partly Linear Model.” *Journal of the Royal Statistical Society. Series B (Methodological)* 48 (2): 244–248.

Hidalgo, Javier, Heejun Lee, Jungyoon Lee, and Myung Hwan Seo. 2023. “Minimax Risk in Estimating Kink Threshold and Testing Continuity.” In *Essays in Honor of Joon Y. Park: Econometric Theory,* 45:233–259. Advances in Econometrics. Emerald Group Publishing Limited.

Hidalgo, Javier, Jungyoon Lee, and Myung Hwan Seo. 2019. “Robust inference for threshold regression models.” *Journal of Econometrics* 210 (2): 291–309.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2): 263–291.

Khmaladze, E. V. 1982. “Martingale Approach in the Theory of Goodness-of-Fit Tests.” *Theory of Probability & Its Applications* 26 (2): 240–257.

Komarova, Tatiana, and Javier Hidalgo. 2023. “Testing nonparametric shape restrictions.” *The Annals of Statistics* 51 (6): 2299–2317.

Kostyshak, Scott. 2017. “Non-parametric Testing of U-shaped Relationships.” *Available at SSRN 2905833.*

Lind, Jo Thori, and Halvor Mehlum. 2010. “With or Without U? The Appropriate Test for a U-Shaped Relationship*.” *Oxford Bulletin of Economics and Statistics* 72 (1): 109–118.

Muller, Hans-Georg. 1989. “Adaptive Nonparametric Peak Estimation.” *Annals of Statistics* 17 (3): 1053–1069.

———. 1992. “Change-Points in Nonparametric Regression Analysis.” *Annals of Statistics* 20 (2): 737–761.

Newell, Jay, Ulrike Genschel, and Ni Zhang. 2014. “Media Discontinuance: Modeling the Diffusion “S” Curve to Declines in Media Use.” *Journal of Media Business Studies* 11 (4): 27–50.

Newey, Whitney K. 1997. “Convergence rates and asymptotic normality for series estimators.” *Journal of Econometrics* 79 (1): 147–168.

Newey, Whitney K., and James L. Powell. 2003. “Instrumental Variable Estimation of Nonparametric Models.” *Econometrica* 71 (5): 1565–1578.

Pardo, Angel, Vicente Meneu, and Enric Valor. 2002. “Temperature and seasonality influences on Spanish electricity load.” *Energy Economics* 24 (1): 55–70.

Parzen, Emanuel. 1962. “On Estimation of a Probability Density Function and Mode.” *Annals of Mathematical Statistics* 33 (3): 1065–1076.

Procházková, Jana. 2005. “Derivative of B-spline function.” *25. konference o geometrii a pocitacove grafice.*

Rice, John. 1986. “Convergence rates for partially splined models.” *Statistics & probability letters* 4 (4): 203–208.

Robinson, P. M. 1988. “Root-N-Consistent Semiparametric Regression.” *Econometrica* 56 (4): 931–954.

Rogers, Everett M. 2003. *Diffusion of innovations* [in eng]. 5th. New York, NY [u.a.]: Free Press, August.

Simonsohn, Uri. 2018. “Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions.” *Advances in Methods and Practices in Psychological Science* 1 (4): 538–555.

Speckman, Paul. 1988. “Kernel Smoothing in Partial Linear Models.” *Journal of the Royal Statistical Society. Series B (Methodological)* 50 (3): 413–436.

Stute, Winfried. 1997. “Nonparametric model checks for regression.” *Annals of Statistics* 25 (2): 613–641.

Stute, Winfried, Silke Thies, and Li-Xing Zhu. 1998. “Model checks for regression: an innovation process approach.” *The Annals of Statistics* 26 (5): 1916–1934.

Sutton, John, and Daniel Trefler. 2016. “Capabilities, Wealth, and Trade.” *Journal of Political Economy* 124 (3): 826–878.

Utterback, James M. 1996. *Mastering the dynamics of innovation.* USA: Harvard Business School Press.

Weiman, C.G. 1977. “A study of occupational stressor and the incidence of disease/risk.”

Weiss, Alexander, James E. King, Miho Inoue-Murayama, Tetsuro Matsuzawa, and Andrew J. Oswald. 2012. “Evidence for a midlife crisis in great apes consistent with the U-shape in human well-being.” *Proceedings of the National Academy of Sciences* 109 (49): 19949–19952.

Wunder, Christoph, Andrea Wiencierz, Johannes Schwarze, and Helmut Küchenhoff. 2013. “Well-being over the life span: Semiparametric evidence from british and german longitudinal data” [in English]. *Review of economics and statistics* 95 (1): 154–167.

Appendix

Appendix 3.A Nonparametric vs quadratic fits,

The purpose of this Appendix is to illustrate that the choice of quadratic specifications can be very misleading when one tries to estimate (inverse) U-shaped relations. Here we outline several scenarios.

We use the following setting: $y = m(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $x \sim \mathcal{U}[0, 1]$, and ε is independent of x.

Scenario 1. $m(x) = (x^{1/4} - 0.5)^2$, $\sigma = 0.01$. The switch point for this regression function is $1/16 = 0.0625$, but it is not symmetric around this point. This can be seen in Figure 3.10 which shows one set of generated data (1,000 points) from this model and a fitted line using a quadratic specification.

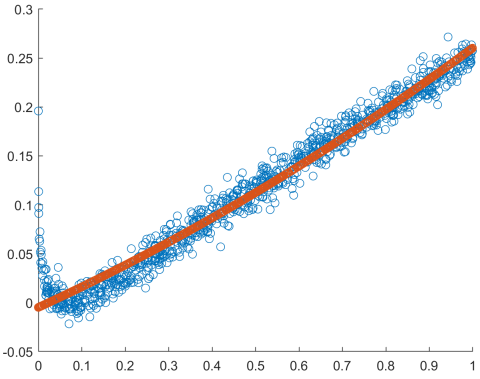
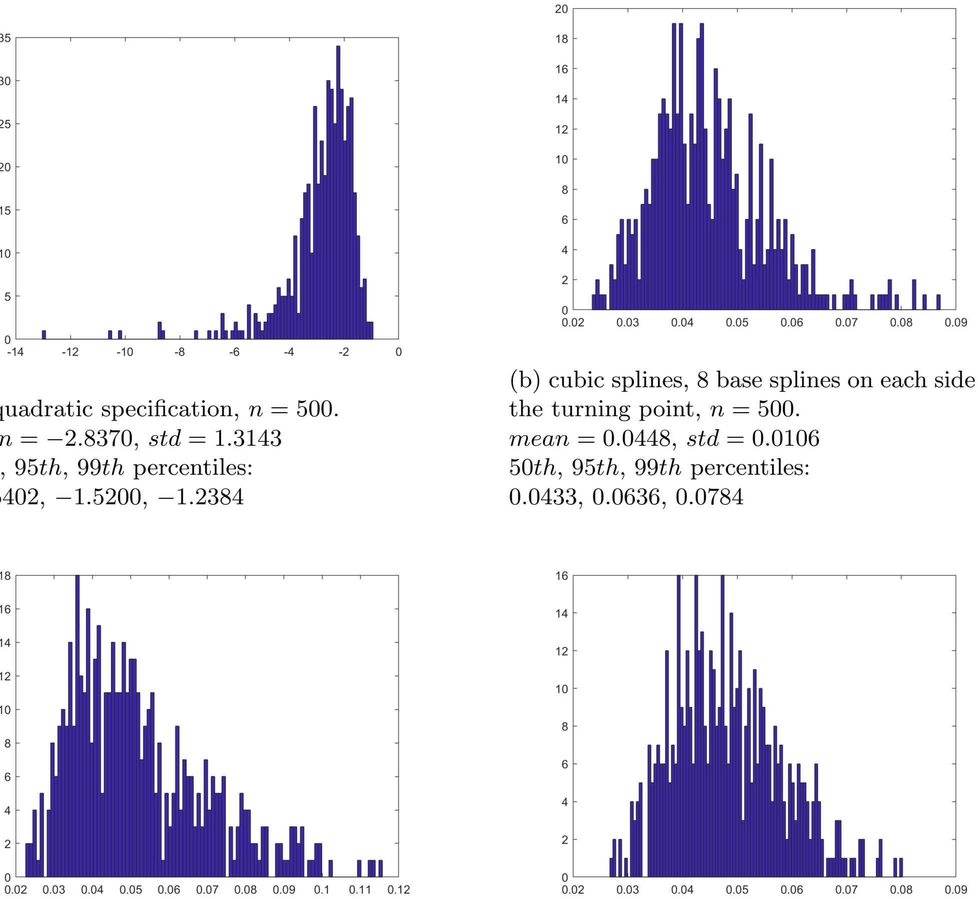


Figure 3.10: Scenario 1.

As can be seen in 3.10, the fitted line is monotonic on the whole domain. Indeed, it turns out that the use of quadratic specification results in the estimated turning point being negative with probability almost 1. Figure 3.11 gives histograms for the estimated turning point in 500 simulations as well the basic summary statistics for those turning points in every subcase (see the caption[ ]s), including the quadratic specification subcase in Panel (a). *B-spline* specifications have a vastly superior performance to quadratic specifications, even though they seem to exhibit a negative finite sample bias when estimating the switch point. This is not



(a) quadratic specification, $n = 500$.
$mean = -2.8370$, $std = 1.3143$
$50th$, $95th$, $99th$ percentiles:
-2.5402, -1.5200, -1.2384

(b) cubic splines, 8 base splines on each side of the turning point, $n = 500$.
$mean = 0.0448$, $std = 0.0106$
$50th$, $95th$, $99th$ percentiles:
0.0433, 0.0636, 0.0784

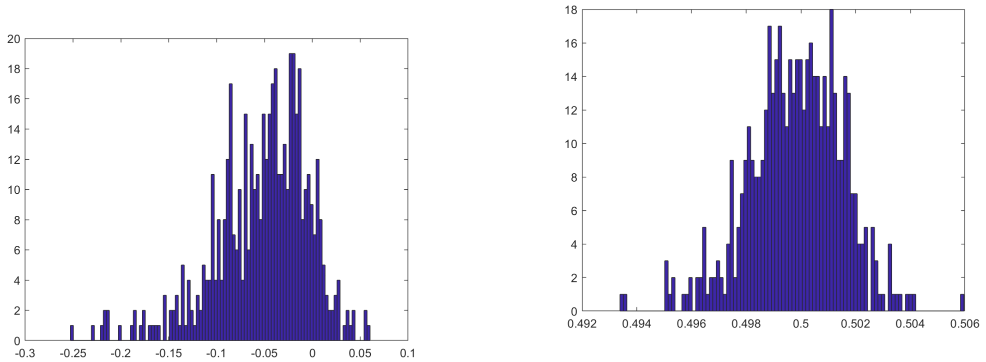
(c) $5th$ degree splines, 13 base splines on each side of the turning point, $n = 2000$
$mean = 0.0522$, $std = 0.0180$
$50th$, $95th$, $99th$ percentiles:
0.0484, 0.0789, 0.0891

(d) cubic splines, 11 splines on each side of the turning point, $n = 2000$
$mean = 0.0482$, $std = 0.01$
$50th$, $95th$, $99th$ percentiles:
0.0473, 0.0651, 0.0744

Figure 3.11: Histograms and summary statistics of estimated switching points in Scenario 1 using various specifications,. Results are obtained in 500 simulations.

surprising given the closeness of the switch point to the boundary. Also, with the sample size increase and the suitable increase in the number of knots the estimated switch points will converge in probability to the true switch point 1/16.

Scenario 2. $m(x) = \Phi(\frac{x-0.5}{5}) \cdot \mathbb{1}(x \leq 0.5) \left(1 - \Phi(\frac{x-0.1}{0.1})\right) \cdot \mathbb{1}(x > 0.5)$, $\sigma = 0.01$. The turning for this regression function is 0.5 and the function is not symmetric around this point. It is continuous at that point but not differentiable (the left and the right derivatives exist and are finite, but they take different values). This can be seen in Figure 3.12 which shows one set of generated data (500 points) from this model and a fitted line using a quadratic specification and a *B-splines* specification with an adaptive choice of a switch point.



(a) quadratic specification, $n = 500$.
$mean = 0.0524$, $std = 0.0497$
$50th$, $95th$, $99th$ percentiles:
-0.0449, 0.0116, 0.0413

(b) cubic splines, 8 base splines on each side of the turning point, $n = 500$.
$mean = 0.4998$, $std = 0.0017$
$50th$, $95th$, $99th$ percentiles:
0.4999, 0.5023, 0.5035

Figure 3.13: Histograms and summary statistics of estimated switching points in Scenario 2 using quadratic and then *B-splines* specifications. Results are obtained in 500 simulations.

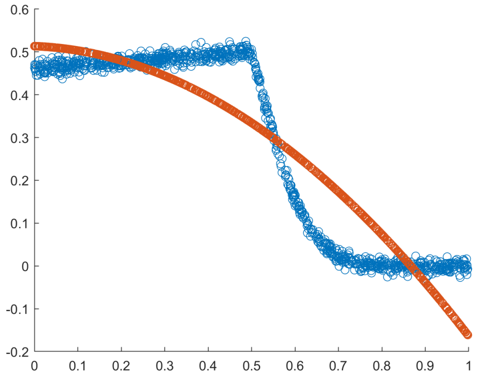


Figure 3.12: Scenario 2.

Figure 3.13 gives histograms for the estimated turning point in 500 simulations as well the basic summary statistics for those turning points in every sub-case (see the caption[ ]s), including the quadratic specification subcase in Panel (a) and the *B-spline* specification in Panel (b). When fitting *B-splines*, we connect two pieces on each side of a turning point as to, first, ensure continuity only (consistent with the property of the original function) and, second, to ensure continuity and differentiability at the switch point. *B-spline* fit also ensures a hump-shaped relation.

Appendix 3.B Proofs

Subsection 3.B.1 Proofs of main results

Proof of Proposition 1.

By Robinson (1988), γ is identified under Condition C3(i), hence the function $m(\cdot)$ is identified as the difference between regression means: $m(x) = E[y|x] - \gamma' E[z|x]$. Suppose, contrary to the statement of the proposition that there are two different ordered sequences $s_1 < s_2 < \ldots s_J$ and $\tilde{s}_1 < \tilde{s}_2 < \ldots \tilde{s}_J$ of switch points such that in addition to (3.6) it holds that

$$m|_{[\tilde{s}_j, \tilde{s}_{j+1}]} \in \mathcal{M}_{j+1}\left([\tilde{s}_j, \tilde{s}_{j+1}]\right), j = 0, \ldots, J, . \tag{3.41}$$

let j_0 be the minimum index such that $s_{j_0} \neq \tilde{s}_{j_0}$. Without a loss of generality, suppose that $s_{j_0} < \tilde{s}_{j_0}$. Using condition (3.7), we then have that on $[s_{j_0}, \tilde{s}_{j_0}]$ the regression function $m(\cdot)$ belongs to $\mathcal{M}_{j_0}\left([s_{j_0}, \tilde{s}_{j_0}]\right)$ (as implied by (3.41)) and also to $\mathcal{M}_{j_0+1}\left([s_{j_0}, \tilde{s}_{j_0}]\right)$ (as implied by (3.6)). But according to (3.8) the intersection $\mathcal{M}_{j_0}\left([s_{j_0}, \tilde{s}_{j_0}]\right) \cap \mathcal{M}_{j_0+1}\left([s_{j_0}, \tilde{s}_{j_0}]\right)$ is empty, which gives us a contradiction. Thus, we can conclude that the ordered sequence of switch points with the properties given in (3.6) is unique. □

Proof of Proposition 2. By Arzela-Ascoli theorem, Θ_0 is relatively compact in the uniform metric. Therefore, its closure $\overline{\Theta}_0$ in the uniform metric is compact. We take $\overline{\Theta}_0$ as our parameter set, and, clearly, $m(x) = E[y|x] - \gamma' E[z|x] \in \overline{\Theta}_0$.

To ensure the compactness of the sample parameter space, as required in the Newey and Powell (2003) (see Section 3.B.3), we use the Arzela-Ascoli theorem once again and obtain the relatively compact set by imposing conditions on the parameters in the *B-spline* approximation captured in the following definition[13] of $\widehat{\Theta}$:

$$\widehat{\Theta} = \left\{ m_{\mathcal{B}} \in \mathcal{M}_{T_{\{(q_j, L_j)\}_{j=1}^{J+1}}} \ : \ |\beta_{\ell_j, j}| \leq A_1 + \Delta_1, \ \frac{L_j |\beta_{\ell_j+1, j} - \beta_{\ell_j, j}|}{s_j - s_{j-1}} \leq A_2 + \Delta_2, \ \forall \ell_j \ \forall j \right\}$$

for some positive constants $\Delta_1 > 0$ and $\Delta_2 > 0$.

As the sample parameter space, we consider the closure $\overline{\widehat{\Theta}}$ of $\widehat{\Theta}$ in the uniform norm. The proof of this proposition establishes, among other things, that every function from $\overline{\Theta}_0$ can be well approximated asymptotically in the uniform metric by functions from $\overline{\widehat{\Theta}}$.

We prove this consistency result by applying Lemma A.1 from Newey and Powell (2003) (see Section 3.B.3). Let us verify all of its conditions. Our population and sample objective functions for the purpose of this proof are, respectively,[14]

$$Q(m(\cdot)) = E[(y - m(x) - \gamma' z)^2], \quad \widehat{Q}\left(m_{\mathcal{B}}(\cdot; s)\right) = \frac{1}{n} \sum_{i=1}^{N} \left(y - m_{\mathcal{B}}(x_i; s) - \gamma' z_i\right)^2 .$$

Condition (i) in Newey and Powell (2003) (Section 3.B.3) about $m(\cdot)$ being the unique

[13] The second condition in this definition is specific to having uniform knots inside each $[s_{j-1}, s_j]$ but could, of course, be easily extended to allow for a different choice of knots.

[14] $\widehat{Q}(\cdot)$ is, of course, $\widehat{Q}^*(\cdot)$ rewritten as a function of the approximation itself.

argmin of Q (up to almost everywhere) in $\overline{\Theta}_0$ follows from the property of the conditional mean as an optimiser and the fact that $m(x) = E[y|x]$ a.e..

For condition (ii) in Newey and Powell (2003) (Section 3.B.3), note that both Q and $\widehat{Q}$ are obviously continuous in m and $m_{\mathcal{B}}$, respectively. Let us show that $\sup_{m\in\overline{\Theta}_0} |Q(m) - \widehat{Q}\,(m)\,| = o_p(1)$. For that, we can use Lemma A.2 in Newey and Powell (2003) (Section 3.B.3) and note that for any $\widetilde{m}, \widetilde{\widetilde{m}} \in \overline{\Theta}_0$

$$\begin{aligned}\left|\widehat{Q}(\widetilde{m}) - \widehat{Q}\left(\widetilde{\widetilde{m}}\right)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{m}(x_i) - \widetilde{\widetilde{m}}(x_i)\right)\left(2m(x_i) + 2u_i - \widetilde{m}(x_i) - \widetilde{\widetilde{m}}(x_i)\right)\right| \\ &\leq \sup_{[\underline{x},\overline{x}]}\|\widetilde{m}(x) - \widetilde{\widetilde{m}}(x)\| \cdot \left(4A_1 + \frac{1}{n}\sum_{i=1}^{n}|u_i|\right),\end{aligned}$$

and of course, $\frac{1}{n}\sum_{i=1}^{n}|u_i| = O_p(1)$ implied by the assumption that u_i has finite fourth moment. Thus, by Lemma A.2 in Newey and Powell (2003) (see Section 3.B.3) we can conclude that

$$\sup_{m\in\overline{\Theta}_0}\left|Q(m) - \widehat{Q}\,(m)\right| = o_p(1). \tag{3.42}$$

Finally, for condition (iii), we want to show that for every $m \in \overline{\Theta}_0$ there is a sequence of $m_{\mathcal{B}} \in \overline{\widehat{\Theta}}$ such that $\sup_x |m_{\mathcal{B}}(x;s) - m(x)| = o(1)$. Note that Condition C2 automatically implies that for every $m \in \Theta_0$ we can find an approximation $m_{\mathcal{B}} \in \mathcal{M}_{T_{\{(q_j,L_j)\}_{j=1}^{J+1}}}$ such that $\sup_x |m_{\mathcal{B}}(x;s) - m(x)| = O\left(\frac{1}{(\min_{j=1,\ldots,J+1} L_j)^r}\right)$ for some $r > 1$, which implies $\sup_x |m_{\mathcal{B}}(x;s) - m(x)| = o(1)$. Let us show that we can take such an approximation $m_{\mathcal{B}}$ to satisfy constraints in the definition of $\widehat{\Theta}$.

First, by *B-spline* properties, $\left|\beta_{\ell_j,j} - \frac{\underline{d}_j+\overline{d}_j}{2}\right| \leq D_{q_j,\infty}\frac{\overline{d}_j-\underline{d}_j}{2}$, where $[\underline{d}_j, \overline{d}_j]$ is the range of values of $m_{\mathcal{B}}(\cdot;s)$ on $[t_{\ell_j+1,j}, t_{\ell_j+q_j-1,j}]$ (see De Boor (1978), p. 133), where $t_{\ell_j,j}$ denotes the ℓ_j's knot on the interval $[s_j, s_{j+1}]$ and $D_{q_j,\infty}$ is a universal constant that does not depend on the system of knots and only depends on the degree of *B-splines* on $[s_j, s_{j+1}]$. Since

$$\begin{aligned}\left|\overline{d}_j - \underline{d}_j\right| &\leq O\left(\frac{1}{L_j^r}\right) + A_2 O\left(\frac{1}{L_j}\right), \\ \left|\overline{d}_j + \underline{d}_j\right| &\leq 2A_1 + O\left(\frac{1}{L_j^r}\right),\end{aligned}$$

then

$$\left|\beta_{\ell_j,j}\right| \leq A_1 + O\left(\frac{1}{L_j}\right) \leq A_1 + \Delta_1$$

for large enough L_j.

Analogously, we can use the same property for the derivative of the *B-spline*. We now have

$$\left|\frac{q_j(\beta_{\ell_j+1,j} - \beta_{\ell_j,j})}{t_{\ell_j+1+q_j,j} - t_{\ell_j+1,j}} - \frac{\underline{e}_j + \overline{e}_j}{2}\right| \le E_{q_j-1,\infty}\frac{\overline{e}_j - \underline{e}_j}{2},$$

where $[\underline{e}_j, \overline{e}_j]$ is the range of values of $m'_{\mathcal{B}}(\cdot)$ on $[t_{\ell_j+1,j}, t_{\ell_j+q_j-2}]$ and $E_{q_j,\infty}$ is a universal constant that does not depend on the system of knots. Once can show that

$$|\overline{c}_j - \underline{c}_j| = O\left(\frac{1}{L_j^{r-1}}\right), \quad |\overline{c}_j + \underline{c}_j| \le 2A_2 + O\left(\frac{1}{L_j^{r-1}}\right).$$

Since $t_{\ell_j+1+q_j,j} - t_{\ell_j+1,j}$ is proportional to $\frac{1}{L_j}$ ($t_{\ell_j+1+q_j,j} - t_{\ell_j+1,j}$ takes possible values of $\frac{s_j-s_{j-1}}{L_j}, 2\frac{s_j-s_{j-1}}{L_j}, \ldots, q_j\frac{s_j-s_{j-1}}{L_j}$), then

$$\frac{L_j|\beta_{\ell_j+1,j} - \beta_{\ell_j,j}|}{s_j - s_{j-1}} \le \left|\frac{q_j(\beta_{\ell_j+1,j} - \beta_{\ell_j,j})}{t_{\ell_j+1+q_j,j} - t_{\ell_j+1,j}}\right| \le A_2 + \Delta_2$$

for large enough L_j.

Now it is only left to consider $m \in \overline{\Theta}_0 \backslash Int(\Theta_0)$, where $Int(\Theta_0)$ denotes the interior of the set Θ_0. For such m we can always find $\tilde{m} \in Int(\Theta_0)$ such that

$$\sup_x |m(x) - \tilde{m}(x)| \le \frac{K_0}{\left(\min_{j=1,\ldots,J+1} L_j\right)^r}$$

for some $K_0 > 0$ (and even a faster rate by the definition of the boundary). Then, according to the discussion above, we can find $m_{\mathcal{B}} \in \overline{\overline{\Theta}}$ such that

$$\sup_x |\tilde{m}(x) - m_{\mathcal{B}}(x; s)| = O\left(\frac{1}{\left(\min_{j=1,\ldots,J+1} L_j\right)^r}\right),$$

implying thus that $\sup_x |m(x) - m_{\mathcal{B}}(x; s)| = O\left(\frac{\tilde{K}_0}{(\min_{j=1,\ldots,J+1} L_j)^r}\right) = o(1)$ as $\min\limits_{j=1,\ldots,J+1} L_j \to \infty$. □

Proof of Corollary 3.5.1. Suppose at least one $\hat{s}_j$ is not consistent for s_j. Let j_0 be the smallest index such that $\hat{s}_{j_0} - s_{j_0} \overset{p}{\nrightarrow} 0$. This means that there is $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ we have $P\left(|\hat{s}_j - s_j| > \varepsilon_1\right) \ge \varepsilon_2$ on a subsequence of $\hat{s}_j$. Without a loss of generality, we can take that $P\left(\hat{s}_{j_0} < s_{j_0} - \varepsilon_1\right) \ge \varepsilon_2$. But we then conclude that on the interval $[s_{j_0} - \varepsilon_1, s_{j_0}]$ the subsequence of $\widehat{m}_{\mathcal{B}}(\cdot)$ with a probability bounded away from zero uniformly approximates the property of the class $\mathcal{M}_{j_0+1}$, which contradicts the fact that the whole sequence $\widehat{m}_{\mathcal{B}}(\cdot)$ on $[s_{j_0} - \varepsilon_1, s_{j_0}]$ converges uniformly in probability to $m(\cdot)$ and on that interval $m(\cdot)$ has property $\mathcal{M}_{j_0}$. Since classes $\mathcal{M}_{j_0}$ and $\mathcal{M}_{j_0+1}$ do not intersect, we obtain a contradiction. Hence, all $\hat{s}_j$ are consistent.
□

Proof of Proposition 3.

The rates of convergence of *B-spline* coefficients and the coefficients of the partial linear model are standard (see e.g. Newey (1997), Robinson (1988)). We focus on the rate of convergence of the switch point estimator.

In order to derive the rates, we split the estimation procedure into two steps: in the first step we fix s and find the parameters $\hat{\beta}(s), \hat{\gamma}(s)$ which minimise the constrained optimisation problem treating the given value of s as a parameter, and in the second step we minimise the redefined objective function $\widehat{Q}(s) = \widehat{Q}^*(s, \hat{\beta}(s), \hat{\gamma}(s))$ with respect to s only. Since $\hat{s}$ minimises $\widehat{Q}(s)$:

$$\left.\frac{\partial \widehat{Q}(s)}{\partial s'}\right|_{s=\hat{s}} = 0 \tag{3.43}$$

and by Taylor expansion around the true s^0:

$$0 = \frac{\partial \widehat{Q}(\hat{s})}{\partial s'} = \frac{\partial \widehat{Q}(s^0)}{\partial s'} + \frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'}(\hat{s} - s^0) + o_p\left(\left|\hat{s} - s^0\right|\right). \tag{3.44}$$

Then:

$$\hat{s} - s^0 = \left(\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} + o_p(1)\right)^{-1} \frac{\partial \widehat{Q}(s^0)}{\partial s'}. \tag{3.45}$$

We show that $\frac{\partial \widehat{Q}(s^0)}{\partial s'}$ is $O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} = \Omega_p\left(\frac{1}{L}\right)$, hence the rate of convergence[15] of $\hat{s}$ to s^0 is $O_p\left(\frac{L}{\sqrt{n}}\right)$.

Let

$$\mathcal{L}(s, \beta, \gamma, \lambda) = \widehat{Q}^*(s, \beta, \gamma) + \lambda g(\beta)$$

be the Lagrangian of the constrained minimisation problem, where the inequality constraints are listed as $g(\beta) \geq 0$ and λ are the corresponding Lagrange multipliers. By the envelope theorem:

$$\frac{\partial \widehat{Q}(s)}{\partial s'} = \frac{\partial \mathcal{L}(\hat{\beta}(s), \hat{\gamma}(s), s, \hat{\lambda}(s), \hat{\mu}(s))}{\partial s'} = \frac{\partial \widehat{Q}^*(\hat{\beta}(s), \hat{\gamma}(s), s)}{\partial s'},$$

so to find the first derivative we only need to differentiate $\widehat{Q}^*(s, \beta, \gamma)$ directly with respect to s

[15] A faster rate of convergence can be achieved in the case where the estimated function has a discontinuity (or discontinuity in a derivative) at the switch point, see e.g.Muller (1992). In applications where the researcher has knowledge of these kinds of changes in behaviour at $\hat{s}$ alternative methods of estimation can be used to find the switch point before estimating the remaining parameters.

and evaluate at $\hat{\beta}(s), \hat{\gamma}(s)$. For any $j \in \{1, 2, \ldots, J\}$:

$$\begin{aligned}
\frac{\partial \widehat{Q}(s^0)}{\partial s_j} &= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - \widehat{m}_{\mathcal{B}}(x_i; s^0) - \hat{\gamma}'\left(s^0\right) z_i\right) \frac{\partial \widehat{m}_{\mathcal{B}}(x_i; s^0)}{\partial s_j} \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n} -2\Bigg(m(x_i) - \widehat{m}_{\mathcal{B}}(x_i; s^0) + \left(\gamma - \hat{\gamma}\left(s^0\right)\right)' z_i + u_i\Bigg) \underbrace{\frac{\partial \widehat{m}_{\mathcal{B}}(x_i; s^0)}{\partial s_j}}_{=O_p(1)}}_{O_p\left(\frac{1}{\sqrt{n}}\right)} \\
&= O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

We now justify the rates listed above.

Using derivation in Lemma 3.B.2 and Lemma 3.B.3, and remembering that due to the envelope theorem we only take the derivative with respect to the *B-spline* basis functions and not with respect to $\hat{\beta}(s)$:

$$\begin{aligned}
\frac{\partial \widehat{m}_{\mathcal{B}}(x; s^0)}{\partial s_j} &= \frac{\partial \widehat{m}_{\mathcal{B}}(x; s^0)}{\partial x}\left(\left(\frac{s^0_{j-1} - x}{s^0_j - s^0_{j-1}}\right)\mathbb{1}\left(x \in [s^0_{j-1}, s^0_j)\right) + \left(\frac{x - s^0_{j+1}}{s^0_{j+1} - s^0_j}\right)\mathbb{1}\left(x \in [s^0_j, s^0_{j+1})\right)\right) \\
&= O_p(1).
\end{aligned}$$

This term is stochastically bounded because the derivative of the spline function with respect to x is bounded (we allow coefficients $\hat{\beta}(s)$ from a space $\widehat{\Theta}$ which imposes a common bound of $A_2 + \Delta_2 < \infty$ on the derivative of $m_{\mathcal{B}}(\cdot, s^0)$ across all n and all possible values of x) and the ratios $\left(\frac{s^0_{j-1} - x}{s^0_j - s^0_{j-1}}\right)\mathbb{1}\left(x \in [s^0_{j-1}, s^0_j)\right)$ and $\left(\frac{x - s^0_{j+1}}{s^0_{j+1} - s^0_j}\right)\mathbb{1}\left(x \in [s^0_j, s^0_{j+1})\right)$ are in $[0, 1]$.

The fact that $\frac{1}{n}\sum_{i=1}^{n} m(x_i) - \widehat{m}_{\mathcal{B}}(x_i; s^0) = O_p\left(\frac{1}{\sqrt{n}}\right)$ is shown in Lemma 3.B.5.

Finally, $\frac{1}{n}\sum_{i=1}^{n}\left(\gamma - \hat{\gamma}\left(s^0\right)\right)' z_i + u_i = O_p\left(\frac{1}{\sqrt{n}}\right)$ by standard results for rates of convergence of the linear part of a partly linear model (e.g. Robinson (1988), this could also be shown directly by the same arguments as in Lemma 3.B.5) and of i.i.d. random variables with bounded second moments (e.g. Lindeberg-Levy CLT).

To find the expression for the second derivative of the objective function we introduce the shorthand notation for the residual:

$$\hat{\varepsilon}_i \equiv y_i - \widehat{m}_{\mathcal{B}}\left(x_i; s\right) - \hat{\gamma}' z_i. \tag{3.46}$$

We have $\widehat{Q}(s) = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ and $\frac{\partial \widehat{Q}(s)}{\partial s'} = \frac{2}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\frac{\partial \hat{\varepsilon}_i}{\partial s'}$, hence

$$\frac{\partial^2 \widehat{Q}(s)}{\partial s \partial s'} = \frac{2}{n}\sum_{i=1}^{n}\frac{\partial \hat{\varepsilon}_i}{\partial s'}\frac{\partial \hat{\varepsilon}_i}{\partial s} + \hat{\varepsilon}_i\frac{\partial^2 \hat{\varepsilon}_i}{\partial s \partial s'}$$

The second term is negligible compared to the first one. The first term captures how much the fit worsens when we use an incorrect switch point. It can be shown that

$$\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} \simeq \Omega_p\left(\frac{1}{L}\right).$$

We impose an incorrect constraint when s is different from s^0. As s approaches s^0, their distance becomes smaller than $\frac{1}{L}$, yet we are still imposing the incorrect constraints based on *B-splines*, which worsens the fit over a region proportional to $\frac{1}{L}$. Within that region the loss of fit is proportional to $|s-s^0|$(see Condition C2): after we take a derivative with respect to s, the loss of fit simplifies to a constant over a region proportional to $\frac{1}{L}$. Hence $\frac{2}{n}\sum_{i=1}^{n}\frac{\partial \hat{\varepsilon}_i}{\partial s'}\frac{\partial \hat{\varepsilon}_i}{\partial s} = \Omega_p\left(\frac{1}{L}\right)$.

Finally, by applying the above results to equation (3.45):

$$|\hat{s}-s^0| = \left(\Omega_p\left(\frac{1}{L}\right)\right)^{-1} O_p\left(\frac{1}{\sqrt{n}}\right)$$

hence

$$|\hat{s}-s^0| = O_p\left(\frac{L}{\sqrt{n}}\right).$$

□

Proof of Proposition 4

This follows from the discussion in the main text, in Section 3.5.2. □

Proof of Proposition 5

We wish to show that, after the transformation, the following term is small (i.e. $o_p\left(\frac{1}{\sqrt{n}}\right)$):

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x)\left(m(x_i) - m_{\mathcal{B}}(x_i;\hat{s})\right).$$

The transformation removes the terms linear in $\frac{\partial m_{\mathcal{B}}(x_i;\hat{s})}{\partial s_k}$.

By Taylor expansion, we have:

$$m_{\mathcal{B}}(x_i;\hat{s}) = m_{\mathcal{B}}(x_i;s^0) + \frac{\partial m_{\mathcal{B}}(x_i;\tilde{s})}{\partial s}(s^0-\hat{s})$$

for $\tilde{s}$ between $\hat{s}$ and s^0 (element-wise). The term inside the average in T_1 can be written as:

$$\begin{aligned}
m(x_i) - m_{\mathcal{B}}(x_i;\hat{s}) &= \underbrace{m(x_i) - m_{\mathcal{B}}(x_i;s^0)}_{=O(L^{-r})} + m_{\mathcal{B}}(x_i;s^0) - m_{\mathcal{B}}(x_i;\hat{s}) \\
&= O\left(L^{-r}\right) + \frac{\partial m_{\mathcal{B}}(x_i;\tilde{s})}{\partial s}(\hat{s}-s^0) \\
&= O\left(L^{-r}\right) + \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}(\hat{s}-s^0) \\
&\quad + \Bigg(\underbrace{\frac{\partial m_{\mathcal{B}}(x_i;\tilde{s})}{\partial s} - \frac{\partial m_{\mathcal{B}}(x_i;\hat{s})}{\partial s}}_{=(\hat{s}-\tilde{s})'\frac{\partial^2 m_{\mathcal{B}}(x_i;\check{s})}{\partial s \partial s'} = O(\|\hat{s}-s^0\|_\infty)} + \underbrace{\frac{\partial m_{\mathcal{B}}(x_i;\hat{s})}{\partial s} - \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}}_{=O\left(L^{-(r-1)}\right)} \Bigg)(\hat{s}-s^0) \\
&= O\left(L^{-r} + L^{-(r-1)}\left\|\hat{s}-s^0\right\|_\infty + \left\|\hat{s}-s^0\right\|_\infty^2\right) + \frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}(\hat{s}-s^0)
\end{aligned}$$

We have used the fact that the best *B-spline* approximation of kth derivative of r-times differentiable function is within $O\left(L^{k-r}\right)$ of the approximated function. We also rely on Taylor expansion (of the *B-spline* itself, as shown above, and of its derivative), where $\check{s}$ is another vector between $\hat{s}$ and s^0 (element-wise).

The final term is linear in $\frac{\partial \widehat{m}_{\mathcal{B}}(x_i;\hat{s})}{\partial s}$ and gets removed by the Khmaladze transformation.

The first part is a common upper bound over all x_i: the second derivative of the *B-spline* with respect to the switch points is bounded over the whole domain of x_i, the bound on the fit is also taken uniformly over the whole domain of x_i. Given Proposition 3 and Condition C3:

$$\begin{aligned}
L^{-r} + L^{-(r-1)}\left\|\hat{s}-s^0\right\|_\infty + \left\|\hat{s}-s^0\right\|_\infty^2 &= O_p\left(L^{-r} + \frac{1}{L^{r-2}\sqrt{n}} + \frac{L^2}{n}\right) \\
&= o_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

These terms are small before the transformation, and the transformation takes the form of a projection which can only make the terms smaller. Hence, the transformed term goes to zero faster than $\frac{1}{\sqrt{n}}$. □

Proof of Proposition 6.

The proof follows the same steps as the proof of Theorem 1 in Komarova and Hidalgo (2023) and is omitted. The only differences are

1. We add regressors of the form z_k and $\frac{\partial \widehat{m}_{\mathcal{B}}(x_k;\hat{s})}{\partial s_l}$. The first type takes non-zero[16] values over the whole domain, the second over $(\hat{s}_{l-1}, \hat{s}_{l+1})$. Both of these regions remain bounded away from zero as sample size increases (unlike the basis functions which have support proportional to $\frac{1}{L}$ that goes to zero, causing issues with eigenvalues of the matrix we use

[16] The only case in which the term could be zero is if the true function m is flat within the domain. In this case we can omit the derivative in switch point, or we can introduce trimming of this term if its realised size is too small. Note that the estimator of the derivative with respect to the switch point is consistent, and if the true underlying value is zero we do not need to remove this term and it can be omitted from the transformation.

in the transformation). This addition does not cause any complications.

2. We use regressors based on the estimates $\hat{\beta}, \hat{s}$ derived from the whole sample. This is again not an issue because they are consistent for the true values β_0, s^0: we can show that the limiting behaviour is the same when we use estimates as if we used the true values.

□

Proof of Theorem 3.5.1.

The first statement follows straight from Proposition 4-6 and continuous mapping theorem. The second statement follows by the same arguments as in Proposition 1 in Komarova and Hidalgo (2023). □

Proof of Theorem 3.5.2. This follows by the same arguments as in Theorem 4 in Komarova and Hidalgo (2023) hence the proof is omitted. □

Subsection 3.B.2 Proofs of supporting results

Lemma 3.B.1. B-splines *are continuous in the switch point almost everywhere.*

Proof. We want to analyse continuity of P_i in s. Each element of the vector P_i is of the form $p_{\ell,L_j,[s_{j-1},s_j],q}(x_i)$. The value of $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ does not depend on s_k for $k \notin \{j-1, j\}$, hence $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ is continuous in s_k for $k \notin \{j-1, j\}$.

To show continuity of $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ in s_{j-1} at $s_{j-1} = s$ we want to show that for almost all $x \in [0,1]$: $\lim_{\tilde{s}\longrightarrow s} p_{\ell,L_j,[\tilde{s},s_j],q}(x) = p_{\ell,L_j,[s,s_j],q}(x)$. We use the fact that *B-splines* are invariant under a translation and scaling of the knot sequence (see the result from e.g. Lyche, Manni, and Speleers (2017) restated in Lemma 3.B.7). $p_{\ell,L_j,[s_{j-1},s_j],q}(x_i)$ is defined on the knot sequence

$$t^{[s_{j-1},s_j],L'_j,q} = \left(\underbrace{s_{j-1},\dots,s_{j-1}}_{q+1 \text{ times}}, s_{j-1} + \frac{s_j - s_{j-1}}{L'_j}, s_{j-1} + 2\frac{s_j - s_{j-1}}{L'_j}, \dots, \underbrace{s_j,\dots,s_j}_{q+1 \text{ times}}\right).$$

and moving from $s_{j-1} = \tilde{s}$ to $s_{j-1} = s$ is equivalent to scaling by $\frac{s_j - s}{s_j - \tilde{s}}$ and shifting by $-\frac{(\tilde{s}-s)s_j}{s_j-\tilde{s}}$: $t^{[s,s_j],L'_j,q} = \left(\frac{s_j-s}{s_j-\tilde{s}}\right) t^{[\tilde{s},s_j],L'_j,q} - \frac{(\tilde{s}-s)s_j}{s_j-\tilde{s}}$. Hence for $x \in [0,1] \setminus \{s, s_j\}$:

$$\lim_{\tilde{s}\longrightarrow s} p_{\ell,L_j,[\tilde{s},s_j],q}(x) = \lim_{\tilde{s}\longrightarrow s} p_{\ell,L_j,[s,s_j],q}\left(\left(\frac{s_j - s}{s_j - \tilde{s}}\right) x - \frac{(\tilde{s}-s)\, s_j}{s_j - \tilde{s}}\right) = p_{\ell,L_j,[s,s_j],q}\,(x)$$

by continuity of $p_{\ell,L_j,[s,s_j],q}\,(x)$ in x on $x \in [0,1] \setminus \{s, s_j\}$ and the fact that regardless of the sequence of $\tilde{s}$ the points $\left(\frac{s_j-s}{s_j-\tilde{s}}\right) x - \frac{(\tilde{s}-s)s_j}{s_j-\tilde{s}}$ will eventually fall in (s, s_j) if $x \in (s, s_j)$ or in $[0, s) \cup (s_j, 1]$ if $x \in [0, s) \cup (s_j, 1]$ (where the *B-spline* is identically equal to zero).

Similarly, for continuity in s_j at $s_j = s$, we have $t^{[s_{j-1},s],L'_j,q} = \left(\frac{s-s_{j-1}}{\tilde{s}-s_{j-1}}\right) t^{[s_{j-1},\tilde{s}],L'_j,q} - \frac{(\tilde{s}-s)s_{j-1}}{\tilde{s}-s_{j-1}}$. Hence for $x \in [0,1] \setminus \{s_{j-1}, s\}$:

$$\lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[s_{j-1},\tilde{s}],q}(x) = \lim_{\tilde{s} \longrightarrow s} p_{\ell,L_j,[s_{j-1},s],q}\left(\left(\frac{s-s_{j-1}}{\tilde{s}-s_{j-1}}\right) x - \frac{(\tilde{s}-s)\, s_{j-1}}{\tilde{s}-s_{j-1}}\right) = p_{\ell,L_j,[s_{j-1},s],q}\,(x)$$

by continuity of $p_{\ell,L_j,[s_{j-1},s],q}\,(x)$ in x on $x \in [0,1] \setminus \{s_{j-1}, s\}$ and the fact that regardless of the sequence of $\tilde{s}$ the points $\left(\frac{s_j - s}{s_j - \tilde{s}}\right) x - \frac{(\tilde{s}-s)s_j}{s_j - \tilde{s}}$ will eventually fall in (s_{j-1}, s) if $x \in (s_{j-1}, s)$ or in $[0, s_{j-1}) \cup (s, 1]$ if $x \in [0, s_{j-1}) \cup (s, 1]$.

We have shown that the individual elements of P_i are continuous in s on almost all x. The only potential points of discontinuity are $x = s_j$, but this is not a problem given that we are interested in $\beta' P_i$ and the β coefficients are constrained to give continuity at $x = s_j$ (only the last *B-spline* on $[s_{j-1}, s_j]$ and the first on $[s_j, s_{j-1}]$ have a discontinuity at $x = s_j$, they both take the value of 1 at that point, and we constrain their corresponding coefficients to be equal: $\beta_{L_j,j} = \beta_{1,j+1}$). □

Lemma 3.B.2. *The first derivative of a* B-spline *basis function* $p_{\ell,L_j,[s_{j-1},s_j],q}(x)$ *with respect to* s_k *is:*

$$\begin{aligned}
&\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial s_k} = \\
&= \frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial x}\left(\left(\frac{s_{k-1}-x}{s_k - s_{k-1}}\right) \mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1}-s_k}\right) \mathbb{1}\left(x \in [s_k, s_{k+1})\right)\right) \qquad (3.47) \\
&= q\left(\frac{p_{\ell,L_j,[s_{j-1},s_j],q-1}(x)}{t_{\ell+q}^{[s_{j-1},s_j],L'_j,q} - t_{\ell}^{[s_{j-1},s_j],L'_j,q}} - \frac{p_{\ell+1,L_j,[s_{j-1},s_j],q-1}(x)}{t_{\ell+q+1}^{[s_{j-1},s_j],L'_j,q} - t_{\ell+1}^{[s_{j-1},s_j],L'_j,q}}\right) \times \\
&\quad \times \left(\left(\frac{s_{k-1}-x}{s_k - s_{k-1}}\right) \mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1}-s_k}\right) \mathbb{1}\left(x \in [s_k, s_{k+1})\right)\right).
\end{aligned}$$

Proof. Let

$$t^{[0,1],K,q} = \left(\underbrace{0,\ldots,0}_{q+1 \text{ times}}, \frac{1}{K}, \frac{2}{K}, \ldots, \underbrace{1,\ldots,1}_{q+1 \text{ times}}\right) \qquad (3.48)$$

be the set of knots on $[0,1]$ with K equally spaced intervals and endpoints repeated $q+1$ times. The degree q *B-splines* defined on this set of knots are $\left\{p_{\ell,K+q,[0,1],q}(x)\right\}_{\ell=1}^{K+q}$. Let us consider a set of knots t which can be written as

$$t^{[0,1],K,q} = \alpha(s)t - \beta(s)$$

with the corresponding set of degree q *B-splines* $\{p_{\ell,t,q}(x)\}_{\ell=1}^{K+q}$. By the invariance of *B-splines* to translation/scaling (see Lemma 3.B.7), for any x in the support of t:

$$p_{\ell,t,q}(x) = p_{\ell,K,[0,1],q}(\alpha(s)x + \beta(s))$$

where by construction $\alpha(s)x + \beta(s)$ is in $[0,1]$, the support of $t^{[0,1],K,q}$. Then for any s_k and any x in the support of t:

$$\begin{aligned}
\frac{\partial p_{\ell,t,q}(x)}{\partial s_k} &= \frac{\partial p_{\ell,K,[0,1],q}\left(\alpha(s)x + \beta(s)\right)}{\partial s_k} \\
&= \frac{\partial p_{\ell,K,[0,1],q}\left(y\right)}{\partial y}\,\Big|_{y=\alpha(s)x+\beta(s)}\;\frac{\partial(\alpha(s)x + \beta(s))}{\partial s_k} \\
&= q\left(\frac{p_{\ell,K,[0,1],q-1}\left(\alpha(s)x + \beta(s)\right)}{t^{[0,1],K,q}_{\ell+q} - t^{[0,1],K,q}_{\ell}} - \frac{p_{\ell+1,K,[0,1],q-1}\left(\alpha(s)x + \beta(s)\right)}{t^{[0,1],K,q}_{\ell+q+1} - t^{[0,1],K,q}_{\ell+1}}\right) \\
&\quad \times \left(\frac{\partial\alpha(s)}{\partial s_k}x + \frac{\partial\beta(s)}{\partial s_k}\right) \\
&= q\left(\frac{p_{\ell,K,[0,1],q-1}\left(\alpha(s)x + \beta(s)\right)}{\alpha(s)t_{\ell+q} + \beta(s) - \alpha(s)t_\ell - \beta(s)} - \frac{p_{\ell+1,K,[0,1],q-1}\left(\alpha(s)x + \beta(s)\right)}{\alpha(s)t_{\ell+q+1} + \beta(s) - \alpha(s)t_{\ell+1} - \beta(s)}\right) \times \\
&\quad \times \left(\frac{\partial\alpha(s)}{\partial s_k}x + \frac{\partial\beta(s)}{\partial s_k}\right) \\
&= q\left(\frac{p_{\ell,t,q-1}\left(x\right)}{t_{\ell+q} - t_\ell} - \frac{p_{\ell+1,t,q-1}\left(x\right)}{t_{\ell+q+1} - t_{\ell+1}}\right)\frac{1}{\alpha(s)}\left(\frac{\partial\alpha(s)}{\partial s_k}x + \frac{\partial\beta(s)}{\partial s_k}\right) \\
&= \frac{\partial p_{\ell,t,q}(x)}{\partial x}\frac{1}{\alpha(s)}\left(\frac{\partial\alpha(s)}{\partial s_k}x + \frac{\partial\beta(s)}{\partial s_k}\right).
\end{aligned}$$

For the set of knots defined on $[s_{j-1}, s_j]$:

$$t^{[s_{j-1},s_j],L'_j,q} = \left(\underbrace{s_{j-1},\ldots,s_{j-1}}_{q+1\text{ times}}, s_{j-1} + \frac{s_j - s_{j-1}}{L'_j}, s_{j-1} + 2\frac{s_j - s_{j-1}}{L'_j}, \ldots, \underbrace{s_j,\ldots,s_j}_{q+1\text{ times}}\right).$$

we can write

$$t^{[0,1],L'_j,q} = \frac{1}{s_j - s_{j-1}}t^{[s_{j-1},s_j],L'_j,q} - \frac{s_{j-1}}{s_j - s_{j-1}},$$

i.e. $\alpha(s) = \frac{1}{s_j - s_{j-1}}$ and $\beta(s) = -\frac{s_{j-1}}{s_j - s_{j-1}}$. Then:

$$\frac{1}{\alpha(s)}\left(\frac{\partial\alpha(s)}{\partial s_k}x + \frac{\partial\beta(s)}{\partial s_k}\right) = \begin{cases} \frac{x - s_j}{s_j - s_{j-1}} & \text{if } s_k = s_{j-1} \\ \frac{s_{j-1} - x}{s_j - s_{j-1}} & \text{if } s_k = s_j \\ 0 & \text{if } s_k \notin \{s_{j-1}, s_j\}. \end{cases}$$

Since for $x \in [s_{k-1}, s_k]$ only the $p_{\ell,L_j,[s_{k-1},s_k],q}(x)$ take non-zero values, for

$x \in [0,1] \setminus \{s_1, s_2, \ldots, s_J\}$:

$$\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial s_k}$$
$$= \frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial x} \left(\left(\frac{s_{k-1} - x}{s_k - s_{k-1}} \right) \mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1} - s_k} \right) \mathbb{1}\left(x \in [s_k, s_{k+1})\right) \right).$$

Note that $\left(\frac{s_{k-1}-x}{s_k-s_{k-1}}\right) \mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x-s_{k+1}}{s_{k+1}-s_k}\right) \mathbb{1}\left(x \in [s_k, s_{k+1})\right)$ is continuous for all $x \in [0,1]$, but due to potential discontinuity in $\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q}(x)}{\partial x}$ at the switch points we need to rule out $x \in \{s_1, s_2, \ldots, s_J\}$.

While the derivatives of specific *B-spline* basis functions may be discontinuous at switch points, this is not a problem in our setting because of the continuity and smoothness constraints which ensure that the derivative with respect to x of the constrained *B-spline* $m_{\mathcal{B}}(x; s) \equiv \sum_{j=1}^{J} \sum_{\ell=1}^{L_j} \beta_{\ell,j} p_{\ell,L_j,[s_{j-1},s_j],q_j}(x)$ at $x = s_k$ is well-defined and continuous. □

Lemma 3.B.3. *$\widehat{Q}(\theta)$ is continuously differentiable.*

Proof. The derivatives with respect to β_ℓ and γ_k are clearly continuous:

$$\frac{\partial \widehat{Q}(\theta)}{\partial \beta_{\ell,j}} = \frac{1}{n} \sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right) p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)$$

$$\frac{\partial \widehat{Q}(\theta)}{\partial \gamma_k} = \frac{1}{n} \sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right) z_{ik}$$

as the first bracket is continuous in β, s (see Lemma 3.B.1) and γ, and $p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)$ and z_{ik} are constant. The derivative with respect to s is a bit more involved, but using Lemma

3.B.2 we can show that it is:

$$
\begin{aligned}
\frac{\partial \widehat{Q}(\theta)}{\partial s_k} &= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\frac{\partial m_{\mathcal{B}}(x_i)}{\partial s_k} \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)}{\partial s_k}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)}{\partial x}\times\right. \\
&\qquad \left.\times \underbrace{\left(\left(\frac{s_{k-1}-x}{s_k - s_{k-1}}\right)\mathbb{1}\left(x \in [s_{k-1}, s_k)\right) + \left(\frac{x - s_{k+1}}{s_{k+1}-s_k}\right)\mathbb{1}\left(x \in [s_k, s_{k+1})\right)\right)}_{\equiv A_{s_k}(x_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\beta_{\ell,j}\frac{\partial p_{\ell,L_j,[s_{j-1},s_j],q_j}(x_i)}{\partial x}A_{s_k}(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} -2\left(y_i - m_{\mathcal{B}}(x_i) - \gamma' z_i\right)\frac{\partial m_{\mathcal{B}}(x_i)}{\partial x}A_{s_k}(x_i).
\end{aligned}
$$

$m_{\mathcal{B}}$ is continuously differentiable in x (by properties of spline functions and by the assumption of smoothness at the minimum), hence $\frac{\partial m_{\mathcal{B}}(x_i)}{\partial x}$ is well-defined for all x_i, and $A_{s_k}(x_i)$ is continuous in both s and $x \in [0,1]$. Hence the derivatives with respect to all inputs are continuous. □

Lemma 3.B.4. *Let x be a k-dimensional vector of random variables. The matrix $E\left(xx'\right)$ is invertible if and only if the elements of x are not linearly dependent, i.e. there does not exist a constant vector $v \in \mathbb{R}^k \setminus \{0\}$ such that $x'v = 0$ a.s..*

Proof. For necessity, suppose $\exists v \in \mathbb{R}^k$ such that $v \neq 0$ and $x'v = 0$ *a.s.*. Then with probability one

$$0 = E\left(xx'v\right) = E\left(xx'\right)v$$

for $v \neq 0$, i.e. $rank\left(E\left(xx'\right)\right) < k$ and $E\left(xx'\right)$ is not invertible.

For sufficiency, suppose $E\left(xx'\right)$ is not invertible. Then there must exist a constant vector $v \in \mathbb{R}^k \setminus \{0\}$ such that $E\left(xx'\right)v = 0$. Then we also have

$$0 = v'0 = v'E\left(xx'\right)v = E\left(v'xx'v\right) = E\left(\left(v'x\right)^2\right)$$

which implies that $v'x = 0$ *a.s.*. □

Lemma 3.B.5. $\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{\mathcal{B}}(x_i; s^0) - m(x_i) = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Proof. This is stated for the case without additional covariates and when we use the true switch point s^0. The argument is identical if we look at the version where instead of $\widehat{m}_{\mathcal{B}}(x_i; s^0)$ we use

$\widehat{m}_{\mathcal{B}}(x_i; s^0) + \widehat{\gamma}' z_i$ and instead of $m(x_i)$ we use $m(x_i) + \gamma' z_i$.

Let P denote the matrix of effective *B-splines* (i.e. after imposing all binding constraints) based on switch points s^0 and evaluated at all points $\{x_i\}_{i=1}^n$. Let m, $m_{\mathcal{B}}\left(s^0\right)$ and $\widehat{m}_{\mathcal{B}}\left(s^0\right)$ denote the vectors of the three functions evaluated at all points $\{x_i\}_{i=1}^n$, and let $\hat{\beta}$ and β_0 be vectors of coefficients such that $\widehat{m}_{\mathcal{B}}\left(s^0\right) = P\hat{\beta}$ and $m_{\mathcal{B}}\left(s^0\right) = P\beta_0$.

The term of interest is:

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{m}_{\mathcal{B}}(x_i; s^0) - m(x_i) = \frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}} - m)$$

where ι is a vector of n 1s.

We use the property[17] that for a scalar random variable X_n:

$$X_n - E(X_n) = O_p\left(\sqrt{V\left(X_n\right)}\right).$$

For $X_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_{\mathcal{B}}(x_i; s^0) - m(x_i)$ we start by looking at the expectation. We firstly find an expression for $\widehat{m}_{\mathcal{B}}\left(s^0\right) - m_{\mathcal{B}}\left(s^0\right)$:

$$\begin{aligned}
\widehat{m}_{\mathcal{B}}\left(s^0\right) = P\hat{\beta} &= P(P'P)^{+}P'(m+u) \\
&= P(P'P)^{+}P'(P\beta_0 + m - P\beta_0 + u) \\
&= P\beta_0 + P(P'P)^{+}P'(m - P\beta_0 + u) \\
&= m_{\mathcal{B}}\left(s^0\right) + P(P'P)^{+}P'(m - m_{\mathcal{B}}\left(s^0\right) + u).
\end{aligned}$$

Each element of the $m - m_{\mathcal{B}}\left(s^0\right)$ vector is bounded above by $\|m(x_i) - m_{\mathcal{B}}(x_i; s^0)\|_\infty = O\left(L^{-r}\right)$, hence we have:

$$\begin{aligned}
\frac{1}{n}\iota'(m_{\mathcal{B}}\left(s^0\right) - m) &= \frac{1}{n}\sum_{i=1}^{n} m_{\mathcal{B}}(x_i; s^0) - m(x_i) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|m(x_i) - m_{\mathcal{B}}(x_i; s^0)\|_\infty \\
&= \|m(x_i) - m_{\mathcal{B}}(x_i; s^0)\|_\infty = O\left(L^{-r}\right).
\end{aligned}$$

[17] This follows from Markov's inequality: $\forall \varepsilon > 0$ there exists $C = C(\varepsilon) > 0$ such that

$$P\left(|X_n| > C\sqrt{E\left(X_n^2\right)}\right) \leq \frac{E\left(X_n^2\right)}{C^2 E\left(X_n^2\right)} = C^{-2} < \varepsilon.$$

We can find an upper bound on the length of the vector $m - m_{\mathcal{B}}\left(s^0\right)$ as:

$$\begin{aligned}\|m - m_{\mathcal{B}}\left(s^0\right)\| &= \sqrt{\sum_{i=1}^{n}(m(x_i) - m_{\mathcal{B}}(x_i;s^0))^2}\\ &\leq \sqrt{n\|m(x_i) - m_{\mathcal{B}}(x_i;s^0)\|_\infty^2}\\ &= \sqrt{n}\|m(x_i) - m_{\mathcal{B}}(x_i;s^0)\|_\infty\end{aligned}$$

It follows that:

$$\begin{aligned}\frac{1}{n}\iota' P(P'P)^{-1}P'(m - m_{\mathcal{B}}\left(s^0\right)) &\leq \frac{1}{n}\|\iota\|\|P(P'P)^{-1}P'(m - m_{\mathcal{B}}\left(s^0\right))\|\\ &\leq \frac{1}{n}\sqrt{n}\|m - m_{\mathcal{B}}\left(s^0\right)\|\\ &\leq \frac{1}{n}\sqrt{n}\sqrt{n}\|m(x_i) - m_{\mathcal{B}}(x_i;s^0)\|_\infty\\ &= \|m(x_i) - m_{\mathcal{B}}(x_i;s^0)\|_\infty = O\left(L^{-r}\right).\end{aligned}$$

The first inequality is by the Cauchy-Schwarz inequality and the second comes from the fact that projecting a matrix can only make it shorter.

Combining all of these facts, we have:

$$\begin{aligned}E\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right) - m)\middle| X\right) &= \frac{1}{n}\iota' E(\widehat{m}_{\mathcal{B}}\left(s^0\right) - m_{\mathcal{B}}\left(s^0\right)|X) + \underbrace{\frac{1}{n}\iota'\left(m_{\mathcal{B}}\left(s^0\right) - m\right)}_{=O(L^{-r})}\\ &= \frac{1}{n}\iota' P(P'P)^{-1}P'(m - m_{\mathcal{B}}\left(s^0\right) + \frac{1}{n}\iota'\underbrace{E(u|X)}_{=0}) + O\left(L^{-r}\right)\\ &= \underbrace{\frac{1}{n}\iota' P(P'P)^{-1}P'(m - m_{\mathcal{B}}\left(s^0\right))}_{=O(L^{-r})} + O\left(L^{-r}\right)\\ &= O\left(L^{-r}\right).\end{aligned}$$

Note that the bound of $2\|m(x_i) - m_{\mathcal{B}}(x_i;s^0)\|_\infty$ does not depend on X, it is the same for all X, hence by the law of iterated expectations we also have:

$$E\left(\frac{1}{n}\iota'(m_{\mathcal{B}}\left(s^0\right) - m)\right) = E\left(E\left(\frac{1}{n}\iota'(m_{\mathcal{B}}\left(s^0\right) - m)\middle| X\right)\right) = O\left(L^{-r}\right).$$

For variance:

$$\begin{aligned}
V\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m)\middle| X\right) &= \frac{1}{n^2}\iota' V\left(P(P'P)^{-1}P'(m-m_{\mathcal{B}}\left(s^0\right)+u)+m_{\mathcal{B}}\left(s^0\right)-m\middle| X\right)\iota \\
&= \frac{1}{n^2}\iota' P(P'P)^{-1}P' V\left(u\middle| X\right)P(P'P)^{-1}P'\iota \\
&= \frac{1}{n^2}\iota' P(P'P)^{-1}P'\sigma^2 I P(P'P)^{-1}P'\iota \\
&= \frac{\sigma^2}{n^2}\underbrace{\iota' P(P'P)^{-1}P'P(P'P)^{-1}P'\iota}_{=\|P(P'P)^{-1}P'\iota\|^2\leq\|\iota\|^2=n} \\
&\leq \frac{\sigma^2}{n}.
\end{aligned}$$

For the second equality we use the fact that $m_{\mathcal{B}}\left(s^0\right)$, m and P are deterministic functions of X. The final inequality comes from the fact that $P(P'P)^{-1}P'$ is a projection matrix, and projecting a vector can only make it shorter.[18] This is again a common bound for any choice of X.

By the law of total variance:

$$\begin{aligned}
V\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m)\right) &= E\left(V\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m)\middle| X\right)\right) \\
&\quad + V\left(E\left(\frac{1}{n}\iota'(\widehat{m}_{\mathcal{B}}\left(s^0\right)-m)\middle| X\right)\right) \\
&= E\left(\frac{\sigma^2}{n}\right)+V\left(O\left(L^{-r}\right)\right) \\
&= O\left(\frac{1}{n}+L^{-2r}\right).
\end{aligned}$$

Finally, using $L^{-r}\prec\frac{1}{\sqrt{n}}$:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{\mathcal{B}}\left(x_i;s^0\right)-m\left(x_i\right)=O_p\left(\frac{1}{\sqrt{n}}+L^{-r}\right)=O_p\left(\frac{1}{\sqrt{n}}\right).$$

□

Lemma 3.B.6.

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(x_i<x)\left(m(x_i)-m_{\mathcal{B}}(x_i;\hat{s})\right)=O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. As in Condition C2(iv), Let $f(x_i,s)=m(x_i)-m_{\mathcal{B}}(x_i,s)$ where $m_{\mathcal{B}}(x_i,s)$ is the best possible *B-spline* approximation to $m(x_i)$ which satisfies the constraints under H_0^B. $f(x_i,s)$ is of order $O\left(L^{-r}\right)$ if $s=s^0$ or x_i is sufficiently far from a misspecified switch point. If $s\neq s^0$ and

[18] In fact, B-splines sum to 1, so the vector if 1s is in the span of P and the projection should leave ι unchanged.

x_i is within a neighbourhood of the misspecified switch point, the $f(x_i, s)$ is separated away from zero and does not go to zero as $n \to \infty$, at least for x_i between the true switch point and the switch point used to impose constraints.[19]

In the proof of Proposition 3 we rely on the Taylor-expansion of the objective function around the true switch point:

$$\begin{aligned}\|\hat{s} - s^0\| &= \left(\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} + o_p(1)\right)^{-1} \frac{\partial \widehat{Q}(s^0)}{\partial s'} \\ &= \left(\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'}\right)^{-1} O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

It can be shown that

$$\frac{\partial^2 \widehat{Q}(s^0)}{\partial s \partial s'} \simeq \frac{1}{n}\sum_{i=1}^{n} \frac{\partial f(x_i, s^0)}{\partial s}\frac{\partial f(x_i, s^0)}{\partial s'} \simeq \int \frac{\partial f(x, s^0)}{\partial s}\frac{\partial f(x, s^0)}{\partial s'} dx \sim \max_k \int \left(\frac{\partial f(x, s^0)}{\partial s_k}\right)^2 dx.$$

At the same time, the term of interest is:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i < x) f(x_i, \hat{s}) &\simeq \int_0^x f(x_i, \hat{s}) dx \\ &\simeq \int_0^x \underbrace{f(x_i, s^0)}_{\sim L^{-r}} + \frac{\partial f(x_i, \hat{s})}{\partial s}(\hat{s} - s^0) dx \\ &\simeq O_p\left(L^{-r}\right) + \int_0^x \frac{\partial f(x_i, \hat{s})}{\partial s} dx \frac{O_p\left(\frac{1}{\sqrt{n}}\right)}{max_k \int \left(\frac{\partial f(x, s^0)}{\partial s_k}\right)^2 dx} \\ &= O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

The last equality is because $\int \left(\frac{\partial f(x,s^0)}{\partial s_k}\right)^2 dx$ and $\int \frac{\partial f(x,s^0)}{\partial s_k} dx$ are proportional to each other (both are $O(1)$ over the same region). And the whole term is close to 0 if x is below the misspecified switch point. □

Subsection 3.B.3 Useful results

Lemma 3.B.7 (B-splines are invariant under a translation and/or scaling of the knot sequence (see e.g. Lyche, Manni, and Speleers (2017)).). *Let $p_{\ell,t,q}(x)$ be the lth* B-spline *function of order q based on the knot vector t evaluated at x, and let $\alpha, \beta \in \mathbb{R}$ with $\alpha \neq 0$. Then*

$$p_{\ell,\alpha t+\beta,q}(\alpha x + \beta) = p_{\ell,t,q}(x). \tag{3.49}$$

[19] e.g. when we use an incorrect switch point and impose a constraint of increasing function when the true one is decreasing, the best we can do is choose a constant function at some level between $m(s)$ and $m(s^0)$.

Lemma NP.A1. (Newey and Powell 2003) based on Gallant (1987): Consistency of an extremum estimator. *Let*

$$\hat{\theta}_n = \arg\min_{\theta \in \widehat{\Theta}} \widehat{Q}(\theta)$$

be an extremum estimator based on a sample of size n and assume that there exists a function $Q(\theta)$ and a set Θ such that:

(i) $Q(\theta)$ has a unique minimum on Θ at θ_0;

(ii) $\widehat{Q}(\theta)$ and $Q(\theta)$ are continuous, Θ is compact, and $\max_{\theta\in\Theta}\left|\widehat{Q}(\theta) - Q(\theta)\right| \xrightarrow{p} 0$;

(iii) $\widehat{\Theta}$ are compact subsets of Θ such that for any $\theta \in \Theta$ there exists $\widehat{\theta} \in \widehat{\Theta}$ such that $\widehat{\theta} \xrightarrow{p} \theta$.

Then

$$\hat{\theta}_n \xrightarrow{p} \theta_0.$$

Lemma NP.A2. (Newey and Powell 2003): Uniform convergence. *If*

(i) Θ is a compact subset of a space with norm $\|\theta\|$;

(ii) $\widehat{Q}(\theta) \xrightarrow{p} Q(\theta)$ for all $\theta \in \Theta$;

(iii) there is a $\delta > 0$ and $B_n = O_p(1)$ such that for all $\theta, \tilde{\theta} \in \Theta$, $|\widehat{Q}(\theta) - \widehat{Q}(\tilde{\theta})| \leq B_n \|\theta - \tilde{\theta}\|^{\delta}$,

then $Q(\theta)$ is continuous and

$$\sup_{\theta\in\Theta} |\widehat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0. \tag{3.50}$$