Econometric Analysis of Heterogeneous Treatment and Network Models

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

submitted by: Florian Sarnetzki

April 2015

Abteilungssprecher: Prof. Dr. Eckhard Janeba
Referent: Prof. Dr. Enno Mammen
Korreferent: Prof. Dr. Markus Frölich

Tag der Verteidigung: 12. Mai 2015

Hiermit erkläre ich, dass die Arbeit selbständig angefertigt und die benutzten Hilfsmittel vollständig und deutlich angegeben sind.

Florian Sarnetzki

Acknowledgments

First and foremost, I am indebted to my advisors Enno Mammen and Markus Frölich. I thank them for their constant support and advice. I have benefited a lot from their deep intuitions in both statistics and econometrics. They gave me the freedom to explore networks on my own and offered help whenever I needed it. Without them this thesis would not have been possible.

I am deeply grateful to my co-author Andreas Dzemski. I really enjoyed working with him on our joint paper. Besides my supervisors, he was the main go-to person when I wanted to discuss econometric problems. His ideas have always been of great use.

I also want to thank Anne Leucht and Stephen Kastoryano who allowed me to be their teaching assistant.

I am grateful for helpful comments from Jasper Haller, Carsten Jentsch, Maria Marchenko, Katharina Momsen, Jan Nimczik, Steffen Reinhold, Peter Scholl, Martin Wahl, Christoph Weiler and Marius Weitz.

I acknowledge my gratitude to the CDSE (Center of Doctoral Studies in Economics) and the University of Mannheim. I learned a lot from the offered course program. Moreover, the amazing research environment with many talented colleagues helped me to sharpen my understanding of economics. I am grateful for comments and questions in various seminars. Financial support from the German-Swiss Research Group FOR 916 and the DFG (Deutsche Forschungsgemeinschaft) is gratefully acknowledged as well. Moreover, I would particularly like to thank the German Economic Association who awarded the second chapter of this thesis with the Reinhard Selten Award.

I thank my parents Anke Ebeling and Bernd Sarnetzki, their partners and my two sisters Eva and Alina for their continuous support.

The years of my thesis work would have not been the same without my friends. I am already looking forward to playing football and discussing about tactics with the Los Duderinhos, going clubbing with the Welpenrudel and hanging out with the Gemeinschaft. Let me also mention my running partner Katharina Momsen, Frauke Schleer and Moritz Lenel who became close friends during my time in Mannheim.

Last but certainly not least, I want to thank my girlfriend Sabine for her love and support.

Table of Contents

1. Introduction

My dissertation investigates commonly used testing and estimation procedures and extends these by taking into account more heterogeneity. In chapter 2, me and my co-author Andreas Dzemski provide a new overidentification test that allows for essential heterogeneity. In chapter 3, I prove weak consistency up to a measure preserving transformation for maximum-likelihood estimation of unobserved latent positions in a Euclidean space just based on observable information of the agent's linking behavior. In chapter 4, I propose a new measure of centrality which exploits the latent space structure and identifies agents who connect clusters.

Chapter 2 is mainly focusing on introducing a novel testing procedure which examines the validity of an instrument. In particular, our test allows the agent's outcome to vary with an unobserved variable which also influences the agent's choice of selecting into treatment. This is sometimes referred to as selection on gains in the literature. Taking an example where the outcome is achieving a high school diploma, one might be interested in the direct effect of a female teenager having a baby during high school on the completion probability. In contrast to the Sargent or Hansen test, we can tolerate teenagers who are less likely to finish school to be more likely to become pregnant. Furthermore, we do not have to assume that this effect is equal across all individuals (or follows a pre-specified parametric form). In chapter 2, we will explicitly discuss this example and use our testing procedure with two instruments which are commonly used in the Economics of health literature.

Chapters 3 and 4 are concerned with networks analysis. In this work, networks are going to be an environment where actors link to each other. A typical data set contains information about which actor is linked to whom. To have an example in mind, one can think of friendships as links between individuals. I assume that actors are more likely to link if they share similar unobserved characteristics which can be summarized by a position in a latent Euclidean space. In chapter 3, I show that maximum likelihood Estimation of these positions has the desirable asymptotic property of weak consistency up to a distance preserving transformation. A similar asymptotic feature has been shown for stochastic block models. In block models, the probability of the formation of a link between two nodes depends on their affiliation to blocks. One either needs very restrictive conditions about pre-knowledge of the estimator or a finite number of blocks for the asymptotic results to hold. Assuming a small number of blocks means that many agents have the same linking behavior. The latent space model allows for much more heterogeneity among the agents. The approach of estimating positions in a latent space is further exploited to provide a new measure of centrality in chapter 4. A cluster is a group of actors who share similar unobserved characteristics which makes them very likely to link. Many other measures of centrality have been introduced. Their aim is to understand how fast information flows through a network depending on which actor got a message first or which actor is informed relatively fast independent of who spreads a rumor. Due to their definition of centrality, these measures usually fail to identify agents that are between clusters. They are likely to denote an agent that is within a cluster to be ”central”. The new measure is novel in the sense that it introduces a way to identify agents who connect clusters. Identifying an agent who is located between clusters is interesting in various settings. For example one might be interested in finding a mediator between opposing groups or detecting an intermediary between markets. This issue is further discussed in simulations in chapter 4. The new measure uses a preliminary estimation step that makes it possible to identify which agents belong to which cluster. One can then easily identify the node which is closest to the midpoint between these clusters. In chapter 4, I show that my measure also has an asymptotic justification.

In the following two sub-chapters, I will give a short review of the literature on testing in a framework with essential heterogeneity and of recent developments in the literature about networks.

1.1. Testing in a framework with essential heterogeneity

In their seminal papers, Imbens and Angrist 1994 and Angrist, Imbens, and Rubin 1996 stress that it is possible to identify a treatment effect for individuals whose binary treatment status is shifted by a change of a binary instrument. In their model, individuals are allowed to select into treatment based on unobserved differing gains. Therefore, a rejection of the Hansen test can not necessarily be interpreted as an indicator of an invalid instrument. Under the assumption of a constant direct effect of the treatment on the outcome, the Hansen test is equivalent to a test that compares the direct effects estimated by two different instruments. Accordingly, a rejection of the Hansen test can also be interpreted as a denial of the assumption of a constant direct effect (see Heckman, Schmierer, and Urzua 2010). There are only a few instrument tests that can be used in a framework as in Imbens and Angrist 1994. Balke and Pearl 1997 identify bounds on the outcome distribution of always-takers and never-takers. Kitagawa 2013 extends their assumptions to a setting with continuous outcomes and uses the bounds to develop a specification test for instrument validity. Under the assumption of no defiers, the bounds indicate that for each level of outcome, there are more individuals which do not select into treatment when the instrument is switched off than there are when the instrument is switched on. This is intuitive, because the first subpopulation should contain always-takers as well as compliers, whereas the latter only contains always-takers. A corresponding bound is provided for the never-takers. He uses a variance-weighted Kolmogorov-Smirnov test statistic. To calculate critical values, he relies on a bootstrap algorithm. A similar path is taken by Huber and Mellace 2014, who look at the corresponding moment inequalities. Therefore, they take into account a subset of the testable implications Kitagawa 2013 investigates. They propose additional bootstrap algorithms to take into account that they look at multiple moment conditions and do not want to get too conservative critical values. A further similar test was introduced by Mourifié and Wan 2014 who use an inference strategy based on Chernozhukov, Lee, and Rosen 2013. A different approach was suggested by Fernandez-Val and Angrist 2013. They assume that the heterogeneity of the direct treatment effect can be fully described by observables. To test instruments across different complier populations, they reweight the corresponding LATEs by a factor which depends on observable characteristics. In contrast to this assumption, we will explicitly model an unobservable variable which may directly influence outcome and the selection into treatment. We augment a standard model by assuming that both a binary and a continuous instrument are available. Under treatment monotonicity a parameter which is closely related to the marginal treatment effect is overidentified. We suggest a test statistic and characterize its asymptotic distribution and behavior under local alternatives. In simulations, we investigate the validity and a finite sample performance of a wild bootstrap procedure.

1.2. Networks

There is a rapidly growing branch of literature on networks in various fields of economics. This section will mostly relate to econometric questions and point out some of the answers which were given in the past. There is a huge amount of literature which investigates the issue of estimating effects of an exogenously given network. The literature on peer effects identifies the effects of the membership to a particular peer-group. Papers on social effects take a more complex view into the networks and try to relate the outcomes of the direct neighbors to the outcomes of an individual. Contrarily to that, my focus in this paragraph is on the formation of one particular network. For overviews of the above literatures see Blume et al. 2010 and Advani and Malde 2014. The econometric literature on network formation can be roughly divided into two strings: papers which are concerned with strategic network formation and papers which do not assume that link formation has a strategic motivation, but is rather mostly driven by homophily.

In the former, authors emphasize the game theoretic structure agents are confronted with. They assume that the observed network is an equilibrium outcome of a game, where agents form links to increase their utility. Therefore, the observed links allow the econometrican to identify factors which drive the particular outcome. The approaches used substantially differ in the games and the equilibrium concepts which are assumed to be in the background. Despite the economically desirable strategic idea of this literature, it is difficult to capture all possible strategies. Thus, all current models rely on meeting processes or finite preference-types assumptions which are unrealistic in reality. Christakis et al. 2010 and Mele 2013 made first approaches to introduce estimateable models. They use a meeting process which forms the network. An assumption that agents only care about the current state when they meet makes it possible to characterize the network formation process as a Markov process. By assuming some symmetries in preferences, Mele 2013 is able to summarize the incentives for any player at any state by a potential function which simplifies the analysis. At the local maxima of the potential function the network is in a Nash-equilibrium. By further assumptions on the meeting process and the idiosyncratic shocks, he can pin down a unique stationary distribution. This distribution serves as a likelihood function and helps to estimate the structural parameters by an approximate version of the exchange algorithm. Paula, Richards-Shubik, and Tamer 2014 set identify structural parameters which drive the network preferences of an agent. They assume that in a pairwise stable network, each agent only cares about a finite number of agents (bounded number of friends and bounded length of indirect links) who are only characterized by a particular network type. This allows them to derive restrictions for the share of these agent types in each preference class (i.e. the set of agents who other agents would benefit from having a link to). Based on these restrictions on allocation parameters, they conclude which structural parameters are feasible.

The second string of literature does not assume the complex game-theoretic background. Neither does it try to identify structural preference parameters.Yet, authors want to explain which actors are more likely to form a link without immediately benefiting from it. An early model is the so called Erdös-Rényi-Gilbert model which either assumes a fixed number of actors who all have the same probability of forming a link or a fixed number of edges where the edge positions are unknown. This literature goes back to the 1950s and was mostly concerned with the number and size of components that result from a particular fixed probability for each link. A substantial step was made by the introduction of the p_1-Model by Holland and Leinhardt 1977. This model examines directed networks (edges between nodes have a direction) and contains several parameters which influence the probability of a particular link. These are a base edge probability, an individual specific effect for incoming and outgoing edges and a mutuality effect for each specific link. Without further assumptions, this model can not be identified. Therefore, several approaches have been made to make maximum-likelihood estimates identifiable. Nevertheless, most of these models lack of consistency results. A more global view on networks was proposed by Frank and Strauss 1986. Instead of modeling the probability of each edge between two actors, their approach deals with the likelihood that a particular network as a whole exists. They try to estimate parameters for particular forms like edges, k-stars or triangles which show up in a network. This model is named Exponential Random Graph Model (ERGM), because of an exponential family assumption for the underlying probability model. A generalization which does not explicitly rely on the above forms, but which allows for arbitrary statistics was introduced by Wasserman and Pattison 1996. This model is also referred to as the p^*-model. It was recently pointed out that many ERGMs become asymptotically equivalent to Erdös-Renyi graphs (see Chatterjee, Diaconis, et al. 2013 and Shalizi, Rinaldo, et al. 2013). Furthermore, Shalizi, Rinaldo, et al. 2013 reveal that ERGMs suffer from the fact that the estimated parameters from a subnetwork do not coincide with those of the whole network if the model contains more complex forms than dyads(edges).

A current standard model for asymptotic results is the stochastic block model (Lorrain and White 1971, Fienberg and Wasserman 1981). The idea of the model is that several actors can be sorted into blocks. The probability of a link does not depend on the involved individuals, but solely on their affiliation to the particular blocks. In recent years, starting with a paper of Bickel and Chen 2009, more and more authors have started investigating the asymptotic behavior of these models. Besides a discussion about identifiability, Bickel and Chen 2009 prove consistency results for maximum likelihood estimation of the block affiliation as well as other methods (Girvan modularity methods). Amini et al. 2013 introduce a pseudo-likelihood algorithm which ignores the dependency structure of an undirected network. Consistency results rely on the assumption that the network is sufficiently sparse. The degree of an agent is the expected number his outgoing edges. Zhao, Levina, and Zhu 2012 add additional parameters which account for different degrees. This is referred to as the degree-corrected stochastic block model. By using the same prove method as Bickel and Chen 2009, they are able to prove consistency as well. Another path to enrich block models with more heterogeneity was taken by Airoldi et al. 2009. In their model, the authors let actors take on different block affiliations in each interaction. These memberships are drawn from a Dirichlet distribution. All of the above papers on block models rely on the assumption that there are finitely many blocks. A first asymptotic result for a growing number of blocks was proven by Choi, Wolfe, and Airoldi 2012. Yet, their result is only informative if a majority of actors is known to be sorted correctly. In block models, parameters besides the community labels of the agents are unknown. Bickel et al. 2013 show consistency and asymptotic normality of maximum likelihood estimation of the parameters which identify the probabilities of the formation of a link depending on the block affiliation and the expected share of members from each block in the whole population.

Besides maximum likelihood estimation, other natural ways to detect block-memberships are clustering methods. A first consistency result in a quite restrictive framework was proven by

Snijders and Nowicki 1997. They assume two blocks and that nodes in one of the two blocks have a higher expected degree. Therefore, they introduce a method which sorts the blocks based on their realized degree and prove its consistency. Nowicki and Snijders 2001 discuss less restrictive conditions without giving consistency results. Rohe, Chatterjee, Yu, et al. 2011 show consistency for spectral clustering and allow for an increasing number of blocks. Spectral clustering uses the eigenvectors of the largest eigenvalues of the normalized graph Laplacian and then sorts the nodes by k-means. To approximate the normalized Laplacian, Rohe, Chatterjee, Yu, et al. 2011 need to make very restrictive assumptions on the expected degree of each node. Therefore, their results are only valid in a setting with very dense graphs. A more flexible generalization of the block model was provided by Hoff, Raftery, and Handcock 2002. They present a model which assumes agents to have a latent position in a Euclidean space. Depending on how close these unobserved positions are to each other, actors have a certain probability of forming a link. The positions can only be identified up to a distance preserving transformation. In their estimates, Hoff, Raftery, and Handcock 2002 rely on Bayesian methods and are therefore not interested in consistency. Building on this model, Handcock, Raftery, and Tantrum 2007 explicitly assume that there is a clustering based on normal distributions in the background. Therefore, the positions are distributed via a multivariate normal distribution. The authors estimate mean and variance of these distributions. Instead of this approach, Hoff 2005 adds random sender and receiver effects. As a consequence of the different extensions, Krivitsky et al. 2009 introduce a model which contains all developments.

A few papers start to make use of the additional structure added by the latent space model. Most importantly, Sussman et al. 2012 derive consistency for a clustering method with a random dot product graph (RDPG) in the background. The used method is similar to spectral clustering, but uses a singular value decomposition of the adjacency matrix, the matrix which summarizes which actors are linked, instead of the Laplacian. Nevertheless, they get the same asymptotic rates as Rohe, Chatterjee, Yu, et al. 2011. Therefore, they also have to rely on the restrictive condition of dense graphs.

Recently, more and more literature is concerned with graphon estimation. A graphon is a symmetric function which explains the probability of forming a link depending on an unknown parameter for each involved actor. This function represents the discrete set of link-probabilities between two actors in an infinite dimensional object. A major difficulty is that there is no natural ordering of the unobserved parameters. Wolfe and Olhede 2013 show that graphon estimation based on block models has a mean square error which goes to zero for some ordering. In some sense, assuming a latent space model postulates a graphon which helps to estimate positions in a meaningful way. Chan and Airoldi 2014 propose a sorting and smoothing algorithm. They first sort nodes on their empirical degree to deal with identifiability problems. Airoldi, Costa, and Chan 2013 assume that they observe several graphs. Thus, it is convenient to sort the nodes in such a way that the corresponding blocks have a similar linking behavior through all graphs.

In economics, it is often beneficial to identify agents who are central and therefore more interesting than other agents. Many approaches which are motivated by different ways of looking at the problem have been made in the past. A simple way to characterize a node as central is to count its outgoing edges. This concept is called degree centrality. A logical next step is to declare agents as central who have very well connected neighbors. This idea was further developed by Katz 1953 who introduced a weighted sum of the neighbor's degree as Katz prestige. A generalization named eigenvector centrality was proposed by Yu and Sompel 1965. It weights the centralities of the neighbors. A related approach is used by Banerjee et al. 2013. They count the degree of an actor along more than one stage. Thus, the authors also look at indirect neighbors. Another path was taken by Freeman 1977 and Anthonisse 1971 who count the number of shortest paths an actor is on. Further other definitions of centrality were introduced by different authors. Jackson 2010 offers a short review.

I impose a latent space structure on a block model with an increasing number of classes in a network and give consistency results for a maximum likelihood estimator. In addition, I provide a new centrality measure which exploits the latent space structure and can identify agents who connect clusters. I prove that the new measure has desirable asymptotic features. Using simulations, I investigate the finite sample performance of the estimator and the centrality measure. Finally, I illustrate the usefulness of the new centrality measure in an application concerning political blogs.

2. Overidentification Test in a Nonparametric Treatment Model with Unobserved Heterogeneity

2.1. Introduction

The canonical treatment effect evaluation problem in Economics can be phrased as the problem of recovering the coefficient β from the outcome equation

$$Y = \alpha + \beta D, \tag{1}$$

where D is a binary indicator of treatment status, and α and β are random coefficients. In latent outcome notation[1], the treatment effect β is commonly written as $\beta = Y^1 - Y^0$. If β is known to be constant then it can be identified by classical instrumental variables methods. In this framework it is straightforward to test the validity of the instruments by classical GMM overidentification tests (Hansen 1982, Sargan 1958). In many applications the more natural assumption is to assume that the treatment effect β is non-constant and correlated with D. Economically this means that individuals differ in their gains from participating in the treatment and that when deciding whether to participate or not individuals take into account possible gains from participation. This setting is often referred to as one of *essential heterogeneity* (Heckman, Urzua, and Vytlacil 2006). It was first considered in the seminal papers by Imbens and Angrist 1994 and Angrist, Imbens, and Rubin 1996. These authors give assumptions under which a binary instrument identifies the average treatment effect for the subpopulation of compliers which they dub the Local Average Treatment Effect (LATE). The compliers are the individuals that respond to a change in the realizations of the binary instrument by changing their participation decision. Different instruments may induce different subpopulations to change their treatment status and therefore estimate different LATEs. Hence, if a GMM overidentification test rejects, this no longer constitutes compelling evidence that one instrument is invalid. Rather, it might as well be interpreted as evidence for a non-constant treatment effect (Heckman, Schmierer, and Urzua 2010).

In this paper we present an instrument test that is valid under essential heterogeneity. A key assumption of Imbens and Angrist 1994, which we maintain as well, is treatment monotonicity. Intuitively, this assumption says that individuals can be ordered by their willingness to participate in the treatment. As we show below, an immediate consequence of the monotonicity assumption is that the propensity score, i.e., the proportion of individuals who participate in the treatment, serves as an index that subsumes all information about observed outcomes that is included in a vector of instruments. We test the null hypothesis that this kind of index sufficiency holds as this is a necessary and testable prerequisite for the intractable hypothesis of instrument validity. More concretely, we assume that a binary and a continuous instrument are available. The purpose of the binary instrument is to split the population into two subpopulations. We test whether observed outcomes conditional on the propensity score are identical in the two subpopulations. The reason why we assume continuity of the second instrument is that this offers a plausible way to argue that the supports of the propensity scores in the two subpopulations overlap.

[1] Latent outcomes are defined in Section 2.2. In general, latent outcomes will be functions of observed covariates. As is common in the literature we keep this dependence implicit.

Our test is related to the test of the validity of the matching approach suggested in Heckman et al. 1996 and Heckman et al. 1998. Their test also exploits index sufficiency under the null hypothesis. Moreover, the role that random assignment to a control group serves in their testing approach is similar to the part that the binary instrument plays in our overidentification result. The testing theory that we develop in this paper translates with slight modifications to the testing problem of Heckman et al. 1996 and Heckman et al. 1998. We hope that it will prove useful in other settings where the null hypothesis imposes some kind of index sufficiency as well.

Our testable restriction in terms of a conditional mean function is closely related to a similar restriction in terms of the Marginal Treatment Effect (MTE, see Heckman and Vytlacil 2005 for a discussion of the MTE). The characterization of the restriction in terms of the MTE, while certainly the less practical one for testing, has a lot of theoretical appeal as it illustrates that our test is based on the overidentification of a structural parameter of the model.

We are not the first to consider the problem of testing instruments in a model with essential heterogeneity. Following previous work by Balke and Pearl 1997, Kitagawa 2013 and Huber and Mellace 2014 consider testing the validity of a discrete instrument in a LATE model. They test inequalities for the densities and the mean of the outcomes for always takers and never takers, i.e. two subpopulations for which treatment status is not affected by the instrument. In stark contrast, our test focuses on the subpopulation which responds to the instrument. Fernandez-Val and Angrist 2013 develop a LATE overidentification test under the additional assumption that the heterogeneity is captured by observed covariates. We do not require such an assumption. Our test lends itself naturally to testing continuous instruments, whereas previous tests can handle continuous instruments only via a discretization.

Our method works if both a binary and a continuous instrument are available. This is the case in many relevant applications. In this paper we apply our method to test the validity of instruments that have been used to investigate the effect of teenage child bearing on high school completion. For another example of an evaluation problem where our method would come to bear consider Carneiro, Heckman, and Vytlacil 2011. They estimate returns to schooling using as instruments a binary indicator of distance to college, tuition fees, as well as continuous measures of local labor market conditions.

Our test reduces to the problem of testing the equality of two nonparametric regression curves. This is a problem with a rich history in the statistical literature (cf., e.g., Hall and Hart 1990; King, Hart, and Wehrly 1991; Delgado 1993; Dette and Neumeyer 2001; Neumeyer and Dette 2003). Our testing problem, however, does not fit directly into any of the frameworks analyzed in the previous literature as it comes with the added complication of generated regressors. We propose a test statistic and quantify the effect of the first stage estimation error on the asymptotic distribution of the test statistic. We find that in order to have good power against local alternatives we have to reduce the nonparametric bias from the first stage estimation. With our particular choice of second stage estimator no further bias reduction is necessary.

We propose a bootstrap procedure to compute critical values. In the context of a treatment model with nonparametrically generated regressors Lee 2013 establishes the validity of a multiplier bootstrap that is based on the first order terms in an asymptotic expansion of the underlying process. We suggest a wild bootstrap procedure that does not rely on first order asymptotics and that is easy to implement in standard software. In exploratory simulations our procedure is faithful to its nominal size in small and medium sized samples.

The paper is structured as follows. Section 2.2 defines our heterogeneous treatment model. In Section 2.3 we give an intuitive overview of our method, state our central overidentification result, discuss nonparametric parameter estimation, and define the test statistic. The asymptotic behavior of our test statistic is discussed in Section 2.4. Our simulations are presented in Section 2.5. In Section 2.6 we apply our approach to real data and study the validity of instruments in the context of teenage child bearing and high school graduation. Section 2.7 concludes.

2.2. Model definition

Our version of a treatment model with unobserved heterogeneity in the spirit of Imbens and Angrist 1994 is owed in large part to Vytlacil 2002. As in Abadie 2003 and Frölich 2007 we assume that our assumptions hold conditional on a set of covariates. We restrict ourselves to covariates that take values in a finite set. Our main overidentification result carries over to more general covariate spaces in a straightforward manner. The purpose of the restriction is exclusively to facilitate estimation by keeping the estimation of infinite dimensional nuisance parameters free of the curse of dimensionality. Without loss of generality assume that we can enumerate all possible covariate configurations by $\{1,\ldots,J^{\max}\}$ and let J denote the covariate configuration of an individual. Treatment status is binary and is denoted by D. The latent outcomes are denoted by Y^0 and Y^1 and $Y = (1-D)Y^0 + DY^1$ denotes the observed outcome. Note that by setting $\alpha = Y^0$ and $\beta = Y^1 - Y^0$ we recover the correlated random effects model from equation (1). Let S denote a continuous random variable and let Z denote a binary random variable. Below, S and Z are required to fulfill certain conditional independence assumptions that render them valid instruments in a heterogeneous treatment model. We observe a sample $(Y_i, D_i, S_i, Z_i, J_i)_{i\leq n}$ from (Y, D, S, Z, J). Treatment status is determined by the threshold crossing decision rule

$$D = 1_{\{r_{Z,J}(S)\geq V\}},$$

with $r_{z,j}$ a function that is bounded between zero and one and V satisfying

$$V \sim U[0,1] \quad \text{and} \quad V \perp\!\!\!\perp (S,Z) \mid J. \tag{I-V}$$

Under this assumption the function $r_{z,j}$ is a propensity score and V can be interpreted as an individual's type reflecting her natural proclivity to select into the treatment group. As pointed out in Vytlacil 2002 the threshold crossing model imposes treatment monotonicity.[2] The assumption that V is uniformly distributed is merely a convenient normalization that allows us to identify $r_{z,j}$. The crucial part of this assumption is that the instruments are jointly independent of the heterogeneity parameter V. This allows us to use the instruments

[2] Consider two types $v_1 \leq v_2$. Under the threshold model v_1 participates if v_2 participates. This is independent of the shape of the propensity score function. In particular, monotonicity of the propensity score function in its parameters is not required.

as a source of variation in treatment participation that is independent of the unobserved types. Furthermore, we assume that for given V, Z and J the latent outcomes are independent of S,

$$Y^d \perp\!\!\!\perp S \mid V, Z, J \quad d = 0, 1. \tag{CI-S}$$

Also, for given V and J the latent outcomes are independent of Z,

$$Y^d \perp\!\!\!\perp Z \mid V, J \quad d = 0, 1. \tag{CI-Z}$$

Intuitively, these assumptions state that once the unobserved type is controlled for, the instruments are uninformative about latent outcomes. Note that we do not place any restrictions on the joint distribution of potential outcomes and V. Economically this means that unobserved characteristics such as personal taste that enter into the decision to participate in the treatment are allowed to be correlated with the latent outcomes. The more commonly assumed instrument condition is

$$(Y^0, Y^1, V) \perp\!\!\!\perp (S, Z) \mid J$$

which implies the conditional independence assumptions stated above. To argue the validity of an instrument it is helpful to split up the instrument condition in a way that allows us to disentangle participation and outcome effects. In our application, for example, assumptions CI-S and CI-Z seem quite plausible. The problematic assumption is to assume that the variation in treatment participation induced by the instrument is independent of the variation that is driven by the unobserved types.

Throughout, we let E_z and $\mathrm{E}_{z,j}$ denote the expectation operator conditional on $Z = z$, and $(J, Z) = (j, z)$, respectively.

2.3. Overidentification test

2.3.1. Testing approach

Before we formally introduce the overidentification test we give a heuristic description of our testing approach. Our test is based on comparing observed outcomes in the $Z = 0$ and $Z = 1$ subpopulations. For a fixed covariate configuration j, Figure 2.3.1 shows hypothetical plots for the propensity scores in the two subpopulations. The ranges of the two functions overlap so that there is an interval of participation probabilities that can be achieved in both subpopulations by manipulating the continuous instrument. The lower and upper bound of this interval are denoted by $x_{L,j}$ and $x_{U,j}$, respectively. Consider the participation probability $x^\star$ lying in this interval. Whenever the participation probability $x^\star$ is observed, all types $V \leq x^\star$ will choose to participate in the treatment and all types $V > x^\star$ will abstain from seeking treatment. In other words, if we observe the same propensity score in two subpopulations, then all types will arrive at identical participation decisions regardless of which subpopulation they are selected into. The participation decision fixes which of the two latent outcomes we observe. Therefore, by fixing the propensity score and comparing observed outcomes between the two subpopulations we are in fact comparing latent outcomes. Under the null hypothesis, latent outcomes behave identically in the two subpopulations since by assumption valid instruments

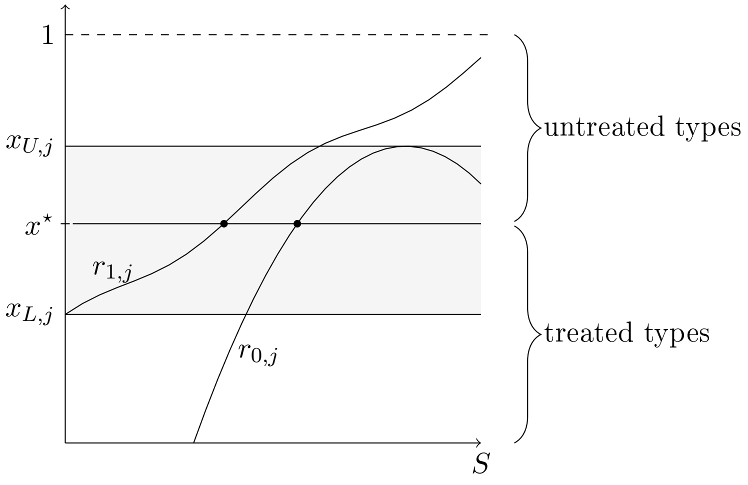


Figure 1: Heuristic description of method.

do not affect latent outcomes. Consequently, for a given propensity score, observed outcomes should behave identically in the $Z = 0$ and $Z = 1$ subpopulations if the model is correctly specified. In particular,

$$\mathrm{E}[Y \mid Z = 0, r_{0,j}(S) = x^\star] = \mathrm{E}[Y \mid Z = 1, r_{1,j}(S) = x^\star].$$

In our approach we test this equality for different $x^\star$.

2.3.2. Overidentification result

For $z = 0, 1$ and $j = 1, \ldots, J^{\max}$ define $m_{z,j}(x) = \mathrm{E}_{z,j}[Y \mid r_{z,j}(S) = x]$. The propensity score is identified from

$$r_{z,j}(s) = \mathrm{E}_{z,j}[D \mid S = s].$$

and therefore $m_{z,j}$ is identified on the interior of the support of $r_{z,j}(S) \mid Z = z$. Our test is based on the following overidentification result.

Proposition 1 (Overidentification) *Fix $j \in \{1, \ldots, J^{max}\}$ and suppose that conditional on $J = j$ x lies in the interior of the support of both $r_{0,j}(S) \mid Z = 0$ and $r_{1,j}(S) \mid Z = 1$. Then $m_{z,j}$ does not depend on z, i.e., $m_{0,j}(x) = m_{1,j}(x)$. Let $m_j(x)$ denote the common value for all j and x that satisfy the assumption.*

Proof

$$\begin{aligned}
m_{z,j}(x) =& \mathrm{E}[Y \mid r_{z,j}(S) = x, Z = z, J = j] \\
=& (1 - x)\,\mathrm{E}[Y^0 \mid r_{z,j}(S) = x, V > x, Z = z, J = j] \\
&+ x\,\mathrm{E}[Y^1 \mid r_{z,j}(S) = x, V \le x, Z = z, J = j] \\
=& (1 - x)\,\mathrm{E}[Y^0 \mid V > x, Z = z, J = j] + x\,\mathrm{E}[Y^1 \mid V \le x, Z = z, J = j] \\
=& (1 - x)\,\mathrm{E}[Y^0 \mid V > x, J = j] + x\,\mathrm{E}[Y^1 \mid V \le x, J = j]
\end{aligned}$$

Now note that the right hand side does not depend on z.

The result says that under the null hypothesis that the model is correctly specified the parameter m_j can be identified from two different subpopulations. Under alternatives the instruments have a direct effect on outcomes that is not mediated through the propensity score. The overidentification restriction has some power to detect such alternatives because in the two subpopulations distinct values of the instrument vector are used to identify the same parameter.

Suppose that for $j = 1, \ldots, J^{\max}$ there are $\underline{x}_j$ and $\bar{x}_j$, $\underline{x}_j \leq \bar{x}_j$, and open sets $\mathcal{G}_j$ such that

$$\operatorname{supp} r_{0,j}(S) \mid Z = 0, J = j \cap \operatorname{supp} r_{1,j}(S) \mid Z = 1, J = j \supseteq \mathcal{G}_j \supseteq [\underline{x}_j \bar{x}_j].$$

Proposition 1 implies that on $[\underline{x}_j \bar{x}_j]$ we have

$$m_{0,j}(x) - m_{1,j}(x) = 0. \tag{2}$$

We are testing this equality. For the test to have some bite we need $[\underline{x}_j \bar{x}_j]$ to be non-empty. Intuitively, what is required is that for fixed Z the continuous instrument is strong enough to induce as many individuals to change their treatment status as would be swayed to change their participation decision by a change in Z while keeping S fixed. An important case where this is not possible is if Z is a deterministic function of S.

The basic idea of the overidentification result does not rely on the continuity of S. However, continuity of S is crucial as it offers a way to ensure that the common support of the propensity scores in the two subpopulations with $Z = 0$ and $Z = 1$ can plausibly have positive probability. For a given j we refer to an interval $[\underline{x}_j, \bar{x}_j]$ that satisfies the above condition as a testable subpopulation. It consists of a set of unobserved types that can be induced to select in and out of treatment by marginal changes in the continuous instrument regardless of the value of the binary instrument. Therefore the types in this interval are part of the complier population as defined in Angrist, Imbens, and Rubin 1996.

Proposition 1 is implied by the stronger result

$$\mathrm{E}_j[Y \mid S, Z] = \mathrm{E}_j[Y \mid r_{Z,j}(S)] \quad a.s.. \tag{3}$$

This says that conditional on covariates, the propensity score aggregates all information that the instruments provide about observed outcomes. In that sense, our approach can be interpreted as a test of index sufficiency that is similar in spirit to the test of the validity of the matching approach suggested in Heckman et al. 1996; Heckman et al. 1998. The equivalence (3) remains true if Y is replaced by a measurable function of Y. By considering different functions of Y a whole host of testable restrictions can be generated. One implication, for example, is that a conditional distribution function is overidentified. In this paper we only consider overidentified conditional mean outcomes and leave the obvious extensions to future research. Our testable restriction (2) is closely related to the marginal treatment effect (MTE)

$$\beta_j(x) = \mathrm{E}_j[Y^1 - Y^0 \mid V = x]$$

which has been proposed as a natural way to parameterize a heterogeneous treatment model (Heckman and Vytlacil 2005). In fact, $\beta_j(x) = \partial_x m_j(x)$. Since we are testing for overidentification of a function, we are also testing for overidentification of its derivative. If we were to base our test directly on the MTE instead of mean outcomes we would not be able to detect alternatives where instruments are uncorrelated with the treatment effect β but have a direct effect on the base outcome α. Another advantage of our mean outcome approach over a test based on the MTE is that we avoid having to estimate a derivative. In our nonparametric setting derivatives are much harder to estimate than conditional means. However, if the econometrician is not interested in a direct effect on the base outcome and if a large sample is available it might be beneficial to look at β_j rather than at m_j. The reason is that as m_j is a smoothed version of β_j it might not provide good evidence for perturbations of β_j that oscillate around zero. Another maybe more compelling reason to consider overidentification of β_j is that it allows us to investigate the source of a rejection of the null hypothesis. If a test based on m_j rejects while at the same time a test based on β_j does not reject it seems likely that instruments have a direct effect on the base outcome but not on the treatment effect. In this paper we focus on the test based on conditional outcomes and leave a test considering the MTE to future research.

It is helpful to think of alternatives as violations of the index sufficiency condition (3). Economically this means that instruments have a direct effect on outcomes, i.e., instruments have an effect on observed outcomes that can not be squared with their role as providers of independent variation in the participation stage. To formalize how our test detects such alternatives ignore covariates for the moment and define the prediction error from regressing on the propensity score instead of on the instruments

$$\varphi(S,Z) = \mathrm{E}[Y \mid S,Z] - \mathrm{E}[Y \mid r_Z(S)].$$

Now suppose that the model is correctly specified up to possibly a violation of the index sufficiency condition. The restricted null hypothesis is

$$H_0 : \varphi(S,Z) = 0 \quad a.s..$$

Using this notation we can rewrite the testable restriction (2) as

$$\mathrm{E}[\varphi(S,Z) \mid r_0(S) = x, Z = 0] - \mathrm{E}[\varphi(S,Z) \mid r_1(S) = x, Z = 1] = 0$$

for all $x \in [\underline{x}, \bar{x}]$. This is a necessary condition for

$$\mathrm{E}[\varphi(S,Z) \mid r_z(S) = x, Z = z] = 0 \quad \text{for } z = 0,1 \text{ and } x \in \operatorname{supp} r_z(S) \mid Z = z$$

which in turn is necessary for the restricted null. Since we are only testing a necessary condition not all alternatives can be detected. As an extreme case consider the case of identical propensity scores, i.e., $r_0 = r_1$. In this particular case our testable restriction does not have the power to detect a direct effect of S on outcomes.

2.3.3. Parameter estimation and test statistic

Let $\hat{m}_{z,j}$ denote an estimator of $m_{z,j}$ and let $\underline{x} = (\underline{x}_1, \dots, \underline{x}_{J^{\max}})$ and $\bar{x} = (\bar{x}_1, \dots, \bar{x}_{J^{\max}})$. Suppose that under the null hypothesis m_j is overidentified on $[\underline{x}_j, \bar{x}_j]$ for $j = 1, \dots, J^{\max}$ and define the test statistic

$$T_n = T_n(\underline{x}, \bar{x}) = \sum_{j=1}^{J} \int_{\underline{x}_j}^{\bar{x}_j} (\hat{m}_{0,j}(x) - \hat{m}_{1,j}(x))^2 \pi_j(x)\, dx. \tag{4}$$

Here π_j is a weight function that can be used to fine-tune power against certain alternatives. What constitutes a sensible choice for π_j will depend on the specifics of the application. For simplicity we assume that π_j is unity from here on. In the following we will refer to the subsample with $J_i = j$ and $Z_i = z$ as the (j, z)-cell. We estimate $\hat{m}_{z,j}$ by a two step procedure. In the first step we estimate the function $r_{z,j}$ by local polynomial regression of D on S within the (j, z)-cell. We will refer to this step as the participation regression. The first step estimator is denoted by $\hat{r}_{z,j}$. In the second step we estimate $m_{z,j}$ by local linear regression of Y on the predicted regressors $\hat{r}_{z,j}(S_i)$ within the (j, z)-cell. This step will be referred to as outcome regression. We let L and K denote the kernel functions for the participation and outcome regression, respectively. Also let g and h denote the respective bandwidth sequences. To reduce notational clutter, we assume that the bandwidths do not depend on j and z. It is straightforward to extend the model to allow cell dependent bandwidths. Let q denote the degree of the local polynomial in the participation regression. It is necessary to choose $q \geq 2$ to remove troublesome bias terms. If these bias terms are not removed the test will behave asymptotically like a linear test, i.e., it will favor the rejection of alternatives that point into a certain direction. A formal definition of the estimators is provided in Appendix A.

In many applications the bounds $\underline{x}$ and $\bar{x}$ are not a priori known and have to be estimated. Below we show that replacing the bounds by a consistent estimator does not affect the asymptotic distribution of the test statistic under weak assumptions. Since we assume $r_{z,j}$ to be continuous, the set on which m_j is overidentified will always be an interval $(x_{L,j}, x_{U,j})$. To avoid boundary problems we fix some positive c_δ and estimate the smaller interval $[\underline{x}_j, \bar{x}_j] = [x_{L,j} + c_\delta, x_{U,j} - c_\delta]$ by its sample equivalent.

2.3.4. Inference and bootstrap method

In Proposition 2 below we characterize the asymptotic distribution of the test statistic under the null. However, as we explain below, we do not recommend to use this distributional result as a basis for approximating critical values. In a related problem with nonparametrically generated regressors Lee 2013 establishes the validity of a multiplier bootstrap procedure. We conjecture that, building on the asymptotic influence function from Lemma 3 in the appendix, a similar approach can be taken in our setting. However, simulating the distribution by multiplier methods has some disadvantages. First, as the approach is based on asymptotic influence functions no improvements beyond first order asymptotics can be expected. Secondly, the method requires significant coding effort which makes it unattractive in applied work. This is why we propose a wild bootstrap procedure that is straightforward to implement instead. We provide simulation evidence that illustrates that the procedure can have good properties in small and medium sized samples. A theoretical proof of the validity of the method is beyond the scope of the present paper and left to future research.

First, estimate the bounds $\underline{x}$ and $\bar{x}$. In the bootstrap samples these bounds can be taken as given. For all j and all z estimate $\hat{r}_{z,j}$ from the (j, z)-cell and predict $R_i^0 = \hat{r}_{Z_i,J_i}(S_i)$ and $\hat{\zeta}_i^0 = D_i - R_i^0$. Next, pool all observations with $J = j$ and estimate m_j by local linear regression of Y_i on R_i^0 with kernel K and bandwidth h. Predict $M_i^0 = \hat{m}_{J_i}(R_i^0)$ and $\hat{\epsilon}_i^0 = Y_i - M_i^0$. Now generate B bootstrap samples in the following way. Draw a sample of n independent

Rademacher random variables $(W_i)_{i\le n}$, let

$$\begin{pmatrix} D_i^* \\ Y_i^* \end{pmatrix} = \begin{pmatrix} R_i^0 \\ M_i^0 \end{pmatrix} + W_i \begin{pmatrix} \hat{\zeta}_i^0 \\ \hat{\epsilon}_i^0 \end{pmatrix},$$

and define the bootstrap sample $(Y_i^*, D_i^*, S_i, Z_i, J_i)_{i\le n}$.

While we use Rademacher variables as an auxiliary distribution, other choices such as the two-point distribution from Mammen 1993 or a standard normal distribution are also possible.

2.4. Asymptotic analysis

In this section we derive the asymptotic distribution of our test statistic. This analysis gives rise to a number of interesting insights. First, it allows us to consider local alternatives. A lesson implicit in the existing literature on L^2-type test statistics is that a naive construction of such a statistic often leads to a test with the undesirable property of treating different local alternatives disparately. Loosely speaking, such a tests behaves like a linear test in that it only looks for alternatives that point to the same direction as a certain bias term (cf. Härdle and Mammen 1993). We find that in order to avoid such behavior it suffices to employ bias-reducing methods when estimating the propensity scores. We recommend to fit a local polynomial of at least quadratic degree. The outcome estimation does not contribute to the problematic bias term. Secondly, our analysis allows us to consider the case when the bounds of integration $\underline{x}$ and $\bar{x}$ are unknown and have to be estimated. We show that, provided that the estimators satisfy a very weak assumption, the asymptotic distribution is unaffected by the estimation. Thirdly, our results allow us to make recommendations about the choice of the smoothing parameters. Our main asymptotic result implies that our test has good power against a large class of local alternatives if the outcome stage estimator oversmoothes compared to the participation stage estimator but not by too much. For convenience of notation, in the following we focus on the case $J^{\max} = 1$ and omit the j subscript. Proofs for the results in this section can be found in the appendix.

2.4.1. Assumptions

Define the sampling errors $\epsilon = Y - \mathrm{E}[Y \mid r_Z(S)]$ and $\zeta = D - E[D \mid S, Z]$. Under the null hypothesis the conditional variances $\sigma_\epsilon^2(x) = \mathrm{E}[\epsilon^2 \mid r_Z(S) = x]$, $\sigma_\zeta^2(x) = \mathrm{E}[\zeta^2 \mid r_Z(S) = x]$ and $\sigma_{\epsilon\zeta}(x) = \mathrm{E}[\epsilon\zeta \mid r_Z(S) = x]$ remain unchanged if the unconditional expectation operator is replaced by the conditional expectation operator E_z, $z = 0, 1$. Also note that $\sigma_\zeta^2(x) = x(1-x)$. For our local estimation approach to work we have to impose some smoothness on the functions m_z and r_z. We now give conditions in terms of the primitives of the model to ensure that the functions that we are estimating are sufficiently smooth.

Assumption 1 *Assume that m is overidentified on an open interval (x_L, x_U) and*

(i) there is a positive ρ such that

$$\mathrm{E}[\exp(\rho |Y^d|)] < \infty, \quad d = 0, 1.$$

(ii) *Conditional on $Z = z$, $z = 0,1$, S is continuously distributed with density $f_{S|Z=z}$ and $r_z(S)$ is continuously distributed with density $f_{R|Z=z}$. Moreover, $f_{S|Z=z}$ is bounded away from zero and has one bounded derivatives and $f_{R|Z=z}$ is bounded away from zero and is twice continuously differentiable.*

(iii) $\mathrm{E}[Y^0 \mid V > x]$ *and* $\mathrm{E}[Y^1 \mid V \le x]$ *are twice continuously differentiable on* (x_L, x_U).

(iv) *The functions* $\mathrm{E}[(Y^0)^2 \mid V > x]$ *and* $\mathrm{E}[(Y^1)^2 \mid V \le x]$ *are continuous on* (x_L, x_U).

(v) r_z, $z = 0,1$, *is* $(q+1)$*-times continuously differentiable on* (x_L, x_U).

The assumption implies standard regularity conditions for m, σ_ϵ^2 and $\sigma_{\epsilon\zeta}$ that are summarized in Assumption 3 in the appendix. These conditions include that m is twice continuously differentiable and that σ_ϵ^2 and $\sigma_{\epsilon\zeta}$ are continuous. A consequence of Assumption 1(ii) is that x_L and x_U are identified by

$$\begin{aligned} x_L &= \max\left\{\inf_s r_0(s), \inf_s r_1(s)\right\} \quad \text{and} \\ x_U &= \min\left\{\sup_s r_0(s), \sup_s r_1(s)\right\}. \end{aligned} \tag{5}$$

Fix a small constant $c_\delta > 0$. We can choose $\underline{x} = x_L + c_\delta$ and $\bar{x} = x_U - c_\delta$. We also need some assumptions about the kernel functions.

Assumption 2 *K and L are symmetric probability density functions with bounded support. K has two bounded and continuous derivatives. The bandwidth sequences are parametrized by $g \sim n^{-\eta^*}$ and $h \sim n^{-\eta}$.*

Implicit in this assumption is that the bandwidths are not allowed to depend on z. In particular, the bandwiths are tied to the overall sample size rather than the size of the two subsamples corresponding to $Z = z$, $z = 0,1$. This is for expositional convenience only.

2.4.2. Local alternatives

To investigate the behavior of the test under local alternatives we now consider a sequence of models that converges to a model in the null hypothesis.

Definition 1 (Local alternative) *A sequence of local alternatives is a sequence of models*

$$\mathcal{M}^n = (Y^{0,n}, Y^{1,n}, V^n, S, Z, r_0, r_1)$$

in the alternative that converges to a model

$$\mathcal{M}^{null} = (Y^{0,null}, Y^{1,null}, V^{null}, S, Z, r_0, r_1)$$

in the null hypothesis in the following sense:

$$\sup_x \mathrm{E}\left[\left(1_{\{V^n \le x\}} - 1_{\{V^{null} \le x\}}\right)^2 \mid S, Z\right] = O_{a.s}\left(c_n^2\right) \tag{6a}$$

$$\mathrm{E}\left[\left(Y^{d,n} - Y^{d,null}\right)^2 \mid S, Z\right] = O_{a.s}\left(c_n^2\right) \quad d = 0,1 \tag{6b}$$

for a vanishing sequence c_n. For n large enough there are positive constants ρ and C such that

$$\mathrm{E}[\exp(\rho |Y^{d,n} - \mathrm{E}[Y^{d,n} \mid S, Z]|) \mid S, Z] \leq C \quad d = 0, 1.$$

We let Y^n and Y^{null} denote the observed outcome under the model $\mathcal{M}^n$ and $\mathcal{M}^{null}$, respectively.

Write φ_n for the index prediction error under the sequence of models $\mathcal{M}^n$ and note that

$$\begin{aligned}\varphi_n(S, Z) &= \mathrm{E}[Y^n \mid S, Z] - \mathrm{E}[Y^n \mid r_Z(S)]\\ &= \mathrm{E}[Y^n - Y^{\mathrm{null}} \mid S, Z] - \mathrm{E}[Y^n - Y^{\mathrm{null}} \mid r_Z(S)] = O_{a.s}(c_n)\end{aligned}$$

so that index sufficiency holds approximately in large samples. Formally, we are testing the sequence of local alternatives

$$H_{0,n} : \Delta_n(x) = 0 \quad \text{for } x \in [\underline{x}, \bar{x}]$$

with

$$\Delta_n(x) = \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x, Z = 0] - \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x, Z = 1].$$

To analyze the behavior of our test under local alternatives we suppose that we are observing a sequence of samples where the n-th sample is drawn from $\mathcal{M}^n$. For vanishing c_n we interpret $\mathcal{M}^{\mathrm{null}}$ as a hypothetical data generating process that satisfies the restriction of the null and that is very close to the observed model $\mathcal{M}^n$. Our objective is to show that our test can distinguish $\mathcal{M}^n$ from $\mathcal{M}^{\mathrm{null}}$. The fastest rate at which local alternatives can be detected is $c_n = n^{-1/2}h^{-1/4}$. This is the standard rate for this type of problem (cf. Härdle and Mammen 1993). At this rate the smoothed and scaled version of the local alternative

$$\Delta_{K,h}(x) = c_n^{-1} \int \Delta_n(x + ht)K(t)\, dt$$

enters the asymptotic distribution of the test statistic.

2.4.3. Asymptotic behavior of the test statistic

For our main asymptotic result below we use the asymptotic framework introduced in the previous subsection where T_n is the test statistic computed on a sample of size n drawn from the model $\mathcal{M}^n$. The result states that the asymptotic distribution of the test statistic can be described by the asymptotic distribution of the statistic under the hypothetical model $\mathcal{M}^{\mathrm{null}}$ shifted by a deterministic sequence that measures the distance of the observed model $\mathcal{M}^n$ from $\mathcal{M}^{\mathrm{null}}$. The behavior of the test statistic under the null is obtained as a special case by choosing a trivial sequence of local alternatives.

Proposition 2 *Let $c_n = n^{-1/2}h^{-1/4}$ and consider a model $\mathcal{M}^{null}$ satisfying Assumption 1 for $x_L < \underline{x} < \bar{x} < x_U$ and corresponding local alternatives $\mathcal{M}^n$ satisfying Definition 1. The functions $\mathrm{E}[Y^n \mid r_Z(S) = x]$ and $\mathrm{E}[Y^n \mid r_Z(S) = x, Z = z]$, $z = 0, 1$, are Riemann integrable on (x_L, x_U). The bandwidth parameters η and η^* satisfy*

$$3\eta + 2\eta^* < 1 \quad (7a) \qquad \eta > {}^1\!/\!_6 \quad (7d)$$
$$2\eta > \eta^* \quad (7b) \qquad (q+1)\eta^* > {}^1\!/\!_2 \quad (7e)$$
$$\eta^* + \eta < {}^1\!/\!_2 \quad (7c) \qquad \eta^* > \eta\,. \quad (7f)$$

Then

$$n\sqrt{h}T_n - \frac{1}{\sqrt{h}}\gamma_n - \int_{\underline{x}}^{\bar{x}} \Delta_{K,h}^2(x)\,dx \xrightarrow{d} N(0,V)\,,$$

where

$$V = 2K^{(4)}(0)\int_{\underline{x}}^{\bar{x}} \left[x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x) + \sigma_\epsilon^2(x)\right]^2 \left(\sum_{z\in\{0,1\}} \frac{1}{p_z f_{R,z}(x)}\right)^2 dx$$

and γ_n is a deterministic sequence such that $\gamma_n \to \gamma$ for

$$\gamma = K^{(2)}(0)\int_{\underline{x}}^{\bar{x}} \left[x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x) + \sigma_\epsilon^2(x)\right] \sum_{z\in\{0,1\}} \frac{1}{p_z f_{R,z}(x)}\,dx.$$

Here $m(x) = E[Y^{null} \mid r_Z(S) = x]$ and the conditional covariances are computed under $\mathcal{M}^{null}$. $K^{(v)}$ denotes the v-fold convolution product of K. For $q \geq 2$ the set of admissible bandwidths is non-empty.

The result implies that the test can detect local alternatives that converge to a model in the null hypothesis at the rate $c_n = n^{-1/2}h^{-1/4}$ and that satisfy

$$\liminf_n \int_{\underline{x}}^{\bar{x}} \Delta_{K,h}^2(x)\,dx > 0.$$

Both the first and the second stage estimation contribute to the asymptotic variance. The term $x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x)$ in the expression for the asymptotic variance is due to the first stage estimation. Under our assumptions this term can not be signed, so that the first stage estimation might increase or decrease the asymptotic variance. However, while it is possible to construct models under which this term is negative, these models have some rather unintuitive features and we do not consider them to be typical. If the estimated regression function is rather flat, the influence of the first stage regression on the asymptotic variance is small. To gain an intuition as to why this is so, note that if $m'(x)$ is small then a large interval of index values around x is informative about $m(x)$. This helps to reduce the first stage estimation error, because on average the index is estimated more reliably over large intervals than over smaller intervals.

An essential ingredient in the proof of Proposition 2 is a result from Mammen, Rothe, and Schienle 2012. They provide a stochastic expansion of a local linear smoother that regresses on generated regressors around the oracle estimator. The oracle estimator is the infeasible estimator that regresses on the true instead of the estimated regressors. This expansion allows us to additively separate the respective contributions of the participation and the outcome regression to the overall bias of our estimator of $m_0 - m_1$. Under the null the oracle estimator is free of bias. This is intuitive. Under the null $m = m_0 = m_1$ so that $\hat{m}_0$ and $\hat{m}_1$ estimate the same function in two subpopulations with non-identical designs. A well-known property of the local linear estimator is that its bias is design independent (Ruppert and Wand 1994) which makes it attractive for testing problems that compare nonparametric fits (Gørgens 2002). Hence, only the bias of the participation regression has to be reduced.

We do not recommend using the distributional result in Proposition 2 to compute critical values. The exact shape of the distribution is very sensitive to bandwidth choice. As explained below, one does not know in practice if bandwidths satisfy the conditions in the theorem. Even if bandwidths are chosen incorrectly, in many cases the statistic still converges to a normal and most of the lessons we draw from the asymptotic analysis still hold up. However, the expressions for the asymptotic bias and variance would look different. Furthermore, to estimate the asymptotic bias and variance we have to estimate derivatives and conditional variances. These are quantities that are notoriously difficult to estimate. Instead, our inference is based on the wild bootstrap procedure introduced in Section 2.3. We investigate the validity of our bootstrap procedure in simulations in Section 2.5 below.

Proposition 2 requires that the bandwidth parameters satisfy a system of inequalities. The restrictions are satisfied for example if $q = 2$, $\eta^* = 1/5$ and $1/6 < \eta < 1/5$. The inequalities (7a)-(7c) ensure that our estimators satisfy the assumptions of Theorem 1 in Mammen, Rothe, and Schienle 2012. Condition (7d) ensures that up to parametric order the bias of the oracle estimator is design independent. When the inequality (7f) is satisfied, the error terms from both the participation and outcome regression contribute to the asymptotic distribution. Finally, inequality (7e) says that the bias from the participation regression must vanish at a faster than parametric rate. This is precisely the condition needed to get rid of the troublesome bias terms discussed above. While the proposition offers conditions on the rates at which the bandwidths should vanish it offers little guidance on how to choose the bandwidths in finite samples. There are no bandwidth selection procedures that produce deliberately under- or oversmoothing bandwidths. This problem is by no means specific to our model but on the contrary quite ubiquitous in the kernel smoothing literature (cf. Hall and Horowitz 2012). In our application we circumvent the problem of bandwidth selection by reporting results for a large range of bandwidth choices.

In practice, the bounds of integration $\underline{x}$ and $\bar{x}$ are additional parameters that have to be chosen. In most applications this means that they have to be estimated from the data. The following result states that a rather slow rate of convergence of these estimated bounds suffices to ensure that bound estimation does not affect the asymptotic distribution.

Proposition 3 *Suppose that the assumptions of Proposition 2 hold. Assume also that $\underline{x}_n$ and $\bar{x}_n$ are sequences of random variables such that*

$$(\underline{x}_n, \bar{x}_n) - (\underline{x}, \bar{x}) = o_p\left(h^{\ell}\right)$$

for a constant $\ell > 1/2$. Then

$$T_n(\underline{x}_n, \bar{x}_n) - T_n(\underline{x}, \bar{x}) = o_p\left(\frac{1}{n\sqrt{h}}\right).$$

alternative	perturbation
1	$\Delta_\alpha = 0.2$
2	$\Delta_\alpha = -\frac{1}{2}V$
3	$\Delta_\alpha = 40(V - 0.3)\exp\left(-80(V - 0.3)^2\right)$
4	$\Delta_\beta = 0.2$
5	$\Delta_\beta = -V$
6	$\Delta_\beta = 40(V - 0.3)\exp\left(-80(V - 0.3)^2\right)$

Tabelle 1: Specification of simulated alternatives.

Let $\hat{x}_L$ and $\hat{x}_U$ denote the sample equivalents of the right hand side of the equation (5) that identifies x_L and x_U, respectively. Under the bandwidth restrictions of Proposition 2 the assumptions in Proposition 3 are satisfied if we set $\underline{x}_n = \hat{x}_L + c_\delta$ and $\bar{x}_n = \hat{x}_U - c_\delta$.

2.5. Simulations

We simulate various versions of the random coefficient model from equation (1) and compute empirical rejection probabilities for our bootstrap test for two sample sizes and a large number of bandwidth choices. As in the previous section we assume $J^{\max} = 1$ and drop the j subscript.

Our basic setup is a model in the null hypothesis. Simulating our test for this model allows us to compare the nominal and empirical size of our test. We then generate several models in the alternative by perturbing outcomes in the basic model for the $Z = 1$ subpopulation. For the basic model we define linear propensity scores $r_0(s) = 0.1 + 0.5s$ and $r_1(s) = 0.5s$. The binary instrument Z is a Bernoulli random variable with $P(Z = 0) = P(Z = 1) = 0.5$ and the continuous instrument S is distributed uniformly on the unit interval. The base outcome α follows a mean-zero normal distribution with variance 0.5. The treatment effect is a deterministic function of V, $\beta = -2V$. As alternatives we consider perturbations of the base outcome α as well as perturbations of the treatment effect β. These perturbations are obtained by adding Δ_α to α and Δ_β to β in the $Z = 1$ subpopulation. The specifications for the alternatives are summarized in Table 1. The first three alternatives consider perturbations of the base outcome, whereas alternatives 4-6 are derived from perturbations of the treatment effect. Alternatives 1 and 4 consider the case that base outcome and treatment effect, respectively, are shifted independently of the unobserved heterogeneity V. The perturbations generating alternatives 2 and 5 are linear functions of V. Finally, alternatives 3 and 6 are generated by perturbing by functions of V that change sign. These alternatives are expected to be particularly hard to detect because our test is based on the m_z function which smoothes over the unobserved heterogeneity as is apparent in the proof of Proposition 1. As bandwidths we choose $g = C_g n^{-\frac{1}{5}}$ and $h = C_h n^{-\frac{1}{6}}$. We report results for a number of choices for the constants C_g and C_h. We set $q = 2$ and choose an Epanechnikov kernel for both K and L. The sample size is set to $n = 200, 400$. These should be considered rather small numbers considering the complexity of the problem. We consider the nominal levels $\theta = 0.1, 0.05$ as these are the most commonly

	$\theta = 0.10$						$\theta = 0.05$					
C_h	0.50	0.75	1.00	1.25	1.50	1.75	0.50	0.75	1.00	1.25	1.50	1.75
null												
$C_g = 0.50$	9.3	8.9	8.4	7.7	8.6	9.6	4.2	3.4	4.7	4.2	4.1	4.1
$C_g = 0.75$	10.1	9.9	9.4	8.2	7.7	9.3	4.8	4.5	4.0	3.3	3.6	4.0
$C_g = 1.00$	8.9	8.7	7.4	9.0	8.9	8.1	4.2	4.1	3.2	4.0	3.6	3.3
alternative 1												
$C_g = 0.50$	94.3	93.8	93.7	93.6	92.8	94.7	88.5	87.1	87.2	86.7	87.7	88.4
$C_g = 0.75$	94.8	91.9	93.0	92.6	94.0	93.8	88.6	86.9	87.2	85.8	87.3	87.2
$C_g = 1.00$	94.0	93.4	94.8	93.5	93.8	93.3	86.7	88.4	89.6	87.2	87.2	89.3
alternative 2												
$C_g = 0.50$	96.9	97.5	97.5	98.1	98.6	98.0	93.3	94.4	95.4	96.0	96.4	95.4
$C_g = 0.75$	96.9	97.9	97.2	97.8	97.1	97.5	93.0	95.6	94.6	94.7	94.3	95.3
$C_g = 1.00$	97.7	97.2	97.4	97.8	97.4	97.8	94.5	95.3	94.1	94.1	95.3	94.4
alternative 3												
$C_g = 0.50$	8.3	8.7	7.2	9.3	8.7	8.9	3.4	3.6	3.5	4.6	4.0	4.0
$C_g = 0.75$	6.9	9.1	8.9	8.6	8.9	9.3	3.5	4.4	3.6	3.6	4.0	3.9
$C_g = 1.00$	8.3	8.2	7.9	8.8	8.9	8.7	4.0	3.7	3.7	3.7	4.6	3.9
alternative 4												
$C_g = 0.50$	25.5	23.8	22.9	24.2	22.6	22.7	15.1	13.9	13.5	13.8	12.3	13.3
$C_g = 0.75$	25.1	26.3	26.1	22.3	23.3	24.7	15.0	14.6	15.0	13.1	13.3	14.5
$C_g = 1.00$	25.4	23.5	24.5	23.7	23.7	23.6	15.2	13.0	15.6	13.8	14.1	13.8
alternative 5												
$C_g = 0.50$	24.3	22.5	21.5	22.7	22.8	21.8	14.9	12.9	11.9	13.8	12.0	12.4
$C_g = 0.75$	21.1	22.0	21.3	20.9	22.7	22.3	10.4	10.8	12.1	11.5	12.7	12.5
$C_g = 1.00$	21.8	21.5	21.4	23.7	21.9	22.5	13.1	12.0	11.3	12.7	12.6	12.2
alternative 6												
$C_g = 0.50$	45.1	44.3	42.3	45.2	46.6	47.7	30.9	30.7	29.3	31.2	35.0	31.2
$C_g = 0.75$	45.3	43.5	44.9	44.2	45.8	44.2	32.3	31.4	32.0	30.7	33.0	30.3
$C_g = 1.00$	44.2	45.5	47.6	44.3	44.6	46.6	32.6	33.7	34.0	30.6	30.9	33.9

Tabelle 2: Empirical rejection probabilities in percentage points under nominal level θ. Sample size is $n = 400$.

used ones in econometric applications. As bound estimation has only a higher order effect we take $\underline{x} = 0.15$ and $\bar{x} = 0.45$ as given. To simulate the bootstrap distribution we are using $B = 999$ bootstrap iterations. For each model we conduct 999 simulations. Empirical rejection probabilities are reported in Table 2 for $n = 400$ and in Table 3 in the appendix for $n = 200$.

We discuss only the results for $n = 400$ in detail. Under the null hypothesis the empirical rejection probabilities are very close to the nominal levels. While this is not conclusive evidence that our bootstrap approach will always work, it is suggestive of the validity of the procedure.

Alternative 1 and Alternative 2 are detected with high probability. These alternatives are particularly easy to detect for two reasons. First, the perturbation affects a large subpopulation so that the alternative is easy to detect due to abundance of relevant data. Secondly, the smoothing inherent in the quantities that our test considers does not smear out the perturbations in a way that makes the alternatives hard to detect. To understand the first effect contrast Alternative 1 and Alternative 2 with Alternative 4 and Alternative 5. Both pairs of alternatives arise from similar perturbations. However, the whole subsample with $Z = 1$ can be used to detect the first pair. In contrast, only treated individuals in the $Z = 1$ subsample provide data that helps to detect the second pair. A back-of-the-envelope calculation reveals that on average only about $400 \times 1/2 \times 1/4 = 25$ observations fall into the subsample with $Z = 1$ and $D = 1$. As cell sizes are observed in applications, a lack of relevant data is a problem that can readily be accounted for when interpreting test results. To shed light on the second effect recall that m_z is derived from smoothing outcomes over $V \leq x$ and $V > x$. Therefore, if a perturbation changes sign, positive and negative deviations from the null will cancel each other out. This effect is precisely what makes it so hard to detect perturbations such as those underlying Alternative 3 and Alternative 6. Luckily, these kinds of alternatives are not what should be expected in many applications. The problem that applied researchers have in mind most of the time is that instruments might have a direct effect on outcomes that can readily be signed by considering the economic context. In that respect, Alternative 1 and Alternative 2 are more typical of issues that applied economists worry about than Alternative 3.

It might seem puzzling that Alternative 6 is detected much more frequently than Alternative 3. The reason is that in Alternative 3 negative deviations in the $V \leq x$ population are offset by positive deviations in the $V > x$ population. This does not happen in Alternative 6 as only the treated population is affected by the perturbation.

Accounting for the complexity of the problem the sample size $n = 200$, for which we report results in the appendix, should be considered very small. Therefore, it is not surprising that the deviations from the nominal size are slightly more pronounced than in the larger sample. The deviations err on the conservative side, but that might be a particularity of our setup. The pattern in the way alternatives are detected is similar to the $n = 400$ sample with an overall lower detection rate.

Our simulations show that our approach has good empirical properties in finite samples. For the simulated model the test holds its size which indicates that the bootstrap procedure works well. Very particular alternatives that perturb outcomes by a function of the unobserved types that oscillates around zero are difficult to detect by our procedure. Alternatives that we consider to be rather typical are reliably detected provided that the subsample affected by the alternative is large enough.

2.6. Application

To illustrate the applicability of our method we now consider the effect of teenage child-bearing on the mother's probability of graduating from high-school. This topic has been discussed extensively in the literature. An early survey can be found in Hoffman 1998. To deal with the obvious endogeneity of motherhood, many authors (Ribar 1994; Hotz, McElroy, and Sanders 2005; Klepinger, Lundberg, and Plotnick 1995) have used instrumental variables methods. It has been suggested that treatment effect heterogeneity is a reason why estimated effects depend strongly on the choice of instrument (Reinhold 2007). In fact, it is very natural to assume that the effect of motherhood on graduation is heterogeneous. For a simple economic model that generates treatment effect heterogeneity suppose that the time cost of child care is the same for students of different abilities whereas the time cost of studying to improve the odds of graduating is decreasing in ability. To translate the problem into our heterogeneous treatment model let D denote a binary indicator of teenage motherhood and let Y denote a binary indicator of whether the woman has obtained a high school diploma[3]. We consider two instruments from the literature. The first one, henceforth labelled S, is age at first menstrual period which has been used in the studies by Ribar 1994 and Klepinger, Lundberg, and Plotnick 1995. This instrument acts as a random shifter of female fecundity and is continuous in nature. Its validity is discussed briefly in Klepinger, Lundberg, and Plotnick 1995 and Levine and Painter 2003. The second instrument, denoted by Z, is an indicator of whether the individual experienced a miscarriage as a teenager. Miscarriage has been used as an unexpected fertility shock in the analysis of adult fertility choices (Miller 2011) and also to study teenage child bearing in Hotz, Mullin, and Sanders 1997; Hotz, McElroy, and Sanders 2005. The population studied in Hotz, McElroy, and Sanders 2005 consists of all women who become pregnant in their teens, whereas we focus on the larger group of all women who are sexually active in their teens. This turns out to be a crucial difference. It stands to investigate the plausibility of the assumptions I-V, CI-S and CI-Z. Arguably, age at first menstrual period is drawn independently of V and fulfills the instrument specific conditional independence assumption CI-S if one controls for race. Possible threats to a linear version of CI-Z are discussed in Hotz, Mullin, and Sanders 1997. Hotz, McElroy, and Sanders 2005 conclude that the linear version of CI-Z holds in good approximation in the population that they are considering. The most problematic assumption to maintain is that Z is orthogonal to V. In a simplified behavioral model teenagers choose to become pregnant based on their unobserved type and then a random draw from nature determines how that pregnancy is resolved. This implies a sort of maximal dependence between Z and V, i.e., teenagers select into treatment and into $Z = 1$ in exactly the same way. Our test substantiates this heuristic argument by rejecting the null hypothesis that the assumptions I-V, CI-S and CI-Z hold simultaneously. Furthermore, it gives instructive insights into the role that heterogeneity plays in the failure of the assumptions.

We use data from the National Longitudinal Survey of Youth 1997[4] (henceforth NLSY97) from round 1 through round 15. We only include respondents who were at least 21 of age at the last interview they participated in. This is to ensure that we capture our outcome variable. A miscarriage is defined as a teenage miscarriage if the woman experiencing the miscarriage was not older than 18 at the time the pregnancy ended. Similarly, a young woman is defined as a teenage mother if she was not older than 18 when the child was born. We control for race for two reasons. First, this is required to make the menarche instrument plausible. Secondly, this takes care of the oversampling of minorities in the NLSY97 so that we are justified in using unweighed estimates. We remove respondents who report “mixed race” as race/ethnicity because the cell size is too small to conduct inference. Table 4 in the appendix gives some summary statistics for our sample. An unfortunate side effect of using the low probability

[3] We do not include equivalency degrees (GED’s). There is a discussion in the literature as to what the appropriate measure is (cf. Hotz, McElroy, and Sanders 2005).

[4] Most of the previous studies relied on data from the National Longitudinal Survey of Youth 1979 (NLSY79). In that study the date of the first menstrual period was asked for for the first time in 1984 when the oldest respondents were 27 years old. As is to be expected, a lot of respondents had trouble recalling the date such a long time after the fact. The NLSY97 contained the relevant question starting from the very first survey when the oldest respondents were still in their teens. Since our method relies on a good measurement of the continuous variable the NLSY97 data is a better choice than the NLSY79 data.

event of a teenage miscarriage as an instrument is that cell sizes can become rather small. This makes it impossible to control for additional covariates while preserving reasonable power. In Section 2.7 we briefly discuss a model that permits a much larger number of covariates. The estimated propensity scores $\hat{r}_{z,j}$ are plotted in Figure 2. For each j the functions $\hat{r}_{0,j}$ and $\hat{r}_{1,j}$ are not identical almost everywhere and their ranges exhibit considerable overlap. We require the same properties from their population counterparts to have good power. It should be noted at this point that the shape of the estimated propensity scores is already indicative of the way that miscarriage fails as an instrument. In a naive telling of the story, the propensity score for women who had a teenage miscarriage is shifted upward, contrary to what we observe in Figure 2. Our test rejects if, keeping the probability of treatment fixed, the difference between the outcomes of the subpopulation with $Z = 0$ and the subpopulation with $Z = 1$ is large. Figure 3 plots $\hat{m}_{0,j}(x) - \hat{m}_{1,j}(x)$ for all values of j. The dashed lines indicate our estimates of $x_{L,j}$ and $x_{U,j}$. We observe that the estimated outcome difference is positive and decreasing

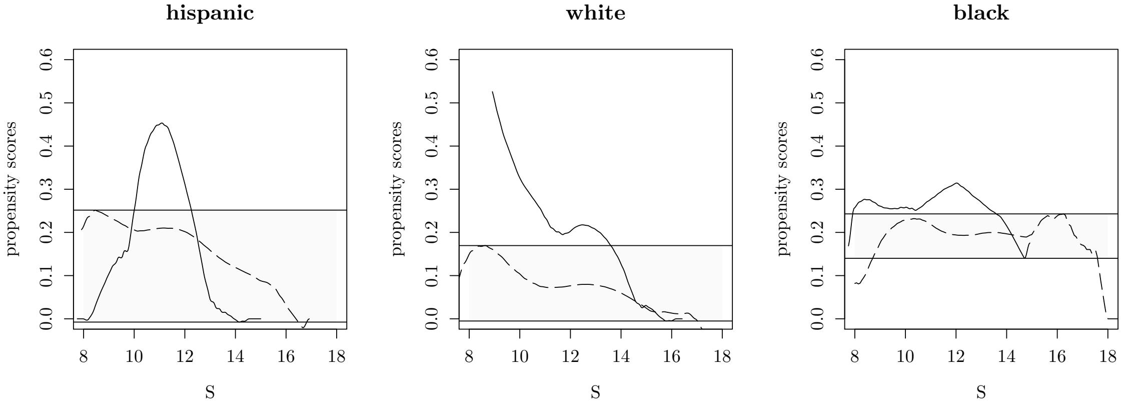


Figure 2: Probability of entering treatment conditional on age of first menstrual period (S) plotted separately for the subpopulations with $Z = 0$ (no miscarriage as a teenager, *dashed line*) and $Z = 1$ (miscarriage as a teenager, *solid line*). Plotted with $q = 1$ and bandwidth $g = 2.00$.

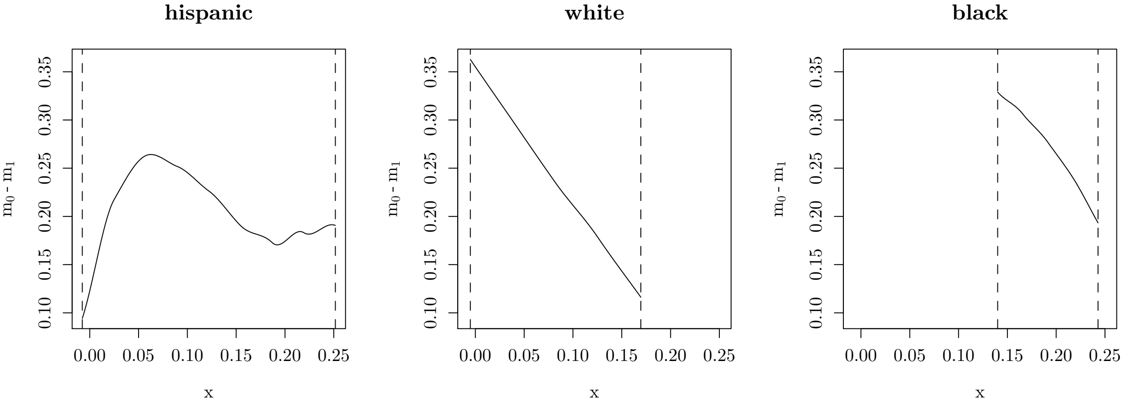


Figure 3: Difference in expected outcomes conditional on probability of treatment between the subpopulations with $Z = 0$ and $Z = 1$. Plotted with $q = 1$ and bandwidths $h = 0.25$ and $g = 2.00$.

in the probability of treatment x. This means that for a low treatment probability x women who have a miscarriage do much worse in terms of high school graduation than do women who do not have a miscarriage. For larger x, however, this difference in outcomes becomes smaller. This feature is in line with our story-based criticism of the instrument. Suppose that the underlying heterogeneity selects women into pregnancy rather than into motherhood. For concreteness think of the heterogeneity as the amount of unprotected sex that a woman has and suppose that this variable is highly correlated with outcomes. In a Bayesian sense a woman who has a miscarriage reveals herself to be of the type that is prone to have unprotected sex. In that sense she is very similar to women with a high probability of becoming pregnant and carrying the child to term and very different from women who become pregnant only with small probability. To turn this eye-balling of the plots in Figure 3 into a rigorous argument we now take into account sampling error by applying our formal testing procedure. For both the first and the second stage regression we choose an Epanechnikov kernel. To have good power against local alternatives we choose $q = 2$. To keep the problem tractable and to reduce the number of parameters we have to choose, we set $g_j = g$ and $h_j = h$ for all j. We then run the test for a large number of bandwidth choices letting h vary between 0.1 and 0.5 and letting g vary between 1 and 3. To determine the bounds of integration $\underline{x}_j$ and $\bar{x}_j$ we use the naive sample equivalence approach suggested in Section 2.4 with different values for c_δ. Table 5 in the appendix reports results for $c_\delta = 0.05$ and Table 6 reports results for $c_\delta = 0.075$. For these two choices of c_δ the test rejects at moderate to high significance levels for a large range of smoothing parameter choices.

Our approach can also be used to investigate other instruments that have been suggested in the literature on teen pregnancies. For example, Z or S could be based on local variation in abortion rates or in availability of fertility related health services (cf. Ribar 1994; Klepinger, Lundberg, and Plotnick 1995).

2.7. Conclusion and Possible Extensions

So far, inference about heterogeneous treatment effect models mostly relies on theoretical considerations about the relationship between instruments and unobserved individual characteristics that are not investigated empirically. This paper shows that under the assumption that a binary and a continuous instrument are available, a parameter is overidentified. This provides a way to test whether the model is correctly specified. The overidentification result is not merely a theoretical curiosity, it has bite when applied to real data. We illustrate this by applying our method to a dataset on teenage child bearing and high school graduation.

Apart from suggesting a new test, we also contribute to the statistical literature by developing testing theory that with slight modifications can be applied to other settings where index sufficiency holds under the null hypothesis. We accommodate an index that is not observed and enters the test statistic as a nonparametrically generated regressor. This setting is encountered, e.g., when testing the validity of the matching approach along the lines suggested in Heckman et al. 1996 and Heckman et al. 1998. Heckman et al. 1998 employ a parametric first-stage estimator. As a result, their second-stage estimator is, to first order, identical to the oracle estimator. Our analysis suggests that replacing the parametric first-stage estimator by a non-

or semiparametric estimator is not innocuous. In particular, it can affect the second-stage bandwidth choice and the behavior of the test under local alternatives.

A theoretical analysis of our wild bootstrap procedure is beyond the scope of this paper. Developing resampling methods for models with nonparametrically generated regressors is an interesting direction for future research. We hope to corroborate the findings in our exploratory simulations by theoretical results in the future.

To apply our method to a particular data set, additional considerations might be necessary. In many applications the validity of an instrument is only plausible provided that a large set of observed variables is controlled for. It is hard to accommodate a rich covariate space in a completely nonparametric model. This is partly due to a curse of dimensionality. Another complicating factor is that our testing approach has good power only if, for fixed covariate values, the instruments provide considerable variation in participation. This is what allows us to test the model for a wide range of unobserved types. Typically, however, instruments become rather weak once the model is endowed with a rich covariate space. These issues can be dealt with by imposing a semiparametric model. As an example, consider the following simple variant of a model suggested in Carneiro and Lee 2009. We let X denote a vector of covariates with possibly continuous components and assume that the unobserved type V is independent of X. Treatment status is determined by $D = 1_{\{R \geq V\}}$ with $R = r_1(X) + r_2(S, Z)$. The unobserved type affects the treatment effect and not the base outcome. The observed outcome is

$$Y = \mu_\alpha(X) + D[\mu_\beta(X) + \lambda(V)].$$

The functions r_1, μ_α and μ_β are known up to a finite dimensional parameter. A semiparametric version of our test would compare $\mathrm{E}[D\lambda(V) \mid R = x, Z]$ in the $Z = 0$ and $Z = 1$ subpopulations. The fact that X is uninformative about V and the additive structure allow for an overidentification result that uses variation in X to extend the interval on which a function is overidentified. This contrasts sharply with Proposition 1 which relies on variation in S keeping the value of covariates fixed. In terms of asymptotic rates this semiparametric model with a large covariate space is not harder to estimate than our fully nonparametric model with a small covariate space and there is no curse of dimensionality.

As seen in Section 2.6 plots of the quantities underlying the test statistic can be helpful in interpreting test results and are a good starting point for discovering the source of a rejection. In many applications it is plausible to assume that while instruments are not valid for extreme types (types with a particularly low or high propensity to participate), they work well for the more average types. The plots can be used to heuristically identify the subpopulation for which instruments are valid. For a subpopulation that based on theoretical considerations is hypothesized to satisfy instrument validity, our approach offers a rigorous way of testing the correct specification of the subpopulation.

3. Consistency of Maximum Likelihood Estimation in a Latent Space Model

3.1. Introduction

The purpose of this chapter is to prove weak consistency for maximum likelihood estimation in a latent space model. In recent years, many authors have studied networks and have given various very different definitions of what they understand as networks. In my case, networks are an environment where actors link to other actors. Data usually consist of nodes that form links and an adjacency matrix that explains whether two nodes are connected or not. Possible examples for networks are individuals who form friendships, firms that collaborate or authors who cite each other. In an application, I examine political internet blogs that hyperlink to each other. A different string of literature is concerned with actors who share a membership to a particular community. Authors studied the effect of this membership (network effects or peer effects). This work takes a step back and investigates whether and how clusters are formed and which individuals end up between these clusters. In order to understand the effects in networks, a fundamental step is to reveal which individuals are likely to form edges and which are not. An other question is which nodes are likely to have a well-balanced linking behavior that make them more interesting in some applications. This is further investigated in chapter 4. Observing more nodes and edges should improve the precision of the answers to the above questions. Therefore, it is important to investigate whether the developed methods have this consistency. I estimate unobserved characteristics of individuals that drive the probability of a link. These characteristics are summarized in a Euclidean space which makes it possible to evaluate which actors are close. I prove weak consistency of maximum likelihood estimation up to a distance preserving transformation under the assumption that the number of possible positions grows. In simulations and in an application about political blogs, I illustrate the applicability of these findings.

Due to the uncommon structure of network data, analysis of asymptotic performance of estimators in network models is a very recent field in statistics. Starting with Bickel and Chen 2009, more and more statistical papers are concerned with networks and their asymptotic behavior. One of the main fields of research about networks relates to community detection. The block model (cf.Wasserman 1994, Lorrain and White 1971, Fienberg and Wasserman 1981) is the current standard model in statistics literature which deals with asymptotic results in networks. In block models, each node is assigned to usually finitely many communities or blocks. The probability of the presence of a link between two nodes solely depends on the community the nodes are assigned to. Consistency of different techniques that estimate this allocation like maximum likelihood among others have recently been proven for different setups with a finite number of blocks (Bickel and Chen 2009, Zhao, Levina, and Zhu 2012). Especially with a small number of classes, block models do not seem to reflect underlying heterogeneity of actors very well. Zhao, Levina, and Zhu 2012 model heterogeneity of nodes within a block through a degree correction. Another approach to allow for more heterogeneity is to model mixed-memberships in a block-model. This path was taken by Airoldi et al. 2009. Nowicki and Snijders 2001 estimate latent characteristics of blocks. They rely on Bayesian methods, have to fix the number of blocks (number of blocks can be unknown in contrast to Bickel and Chen

2009, Zhao, Levina, and Zhu 2012) and do not prove asymptotic results.

In order to allow for more heterogeneity, a promising way is to use block models that admit a growing number of classes. Yet, asymptotic results build on quite restrictive assumptions like linear growing degrees or settings where some pre-knowledge of the estimators outcome has to be assumed. Choi, Wolfe, and Airoldi 2012 allow for a growing number of classes. However, they can only make statements about the number of incorrect class assignments for blocks that already contain a majority of true nodes in their ML-estimate. Therefore, they show a weak consistency result in the spirit of Zhao, Levina, and Zhu 2012 if the majority in each class is known to be specified correctly. To prove my first result, I use similar techniques. By imposing a latent space structure on a block model, I will be able to show weak consistency for a growing number of classes without these restrictive assumptions.

The string of literature concerned with latent space models goes back to Hoff, Raftery, and Handcock 2002. In these models, nodes are assumed to have unobserved positions in a Euclidean space and one aims to find them. Since actors are now allowed to have many different positions, heterogeneity can be modeled much better. The authors use Bayesian techniques to estimate these positions. They form their prior according to maximum likelihood estimation, but do not give further theoretical justification for this step.

Handcock, Raftery, and Tantrum 2007 extend this model by estimating an underlying clustering structure. Among a Bayesian approach, they propose a 2-step ML-estimation procedure, but do not prove asymptotic results. Hoff 2005 adds degree heterogeneity by introducing actor specific random effects. Krivitsky et al. 2009 include all developments in one model. All of these authors use Bayesian estimation.

The above latent space models suffer from the fact that positions in the Euclidean Space can only be identified up to a measure preserving transformation. Having an estimate of the positions that reflects their distances is nevertheless informative regarding which nodes have similar characteristics (or are close in the Euclidean Space) and therefore form clusters and which nodes are located between clusters.

Closely related to my work, Sussman et al. 2012 or Rohe, Chatterjee, Yu, et al. 2011 assume an underlying latent space structure. They deal with clustering algorithms that consistently estimate positions in a stochastic block-model with a number of blocks which increases. Sussman et al. 2012 use spectral clustering which is computationally more efficient than maximum likelihood estimation. Nevertheless, their results rely on settings with many links that seem to be unrealistic in many applications. For example, assuming that the number of friends (or connections) grows at the same rate as individuals in social networks becomes unrealistic for a large number of individuals.

Tang, Sussman, Priebe, et al. 2013 discuss more sparse settings in a framework where they observe a number of vertices tending to infinity. This framework is called semi-supervised in the literature.

This work merges block model literature with latent space literature in the sense that I use methods from Choi, Wolfe, and Airoldi 2012 and Bickel and Chen 2009 to prove a type of consistency of a maximum likelihood estimator concerned with the location in a latent space. In contrast to other papers, I do not rely on semi-supervised settings, assumptions on the realization of the estimator or dense networks. In addition, I use a data set on political blogs that was collected by Adamic and Glance 2005 and check whether a previous sorting into two clusters made by Adamic and Glance 2005 and Zhao, Levina, and Zhu 2012 is reasonable. Proofs to all results are deferred to the appendix.

3.2. Model and a Consistency Result for ML-Estimation in a Latent Space

My version of a latent space model is closely related to Hoff, Raftery, and Handcock 2002. Assume an unobserved characteristic of an agent i can be summarized by z_i^0 which is distributed on K points $\{M_1, \ldots, M_K\} \equiv \mathcal{M} \subset \mathbb{R}^d$. $\mathcal{M}$ is bounded and called latent space. Concentrating on mass points allows me to use proof methods from the block model literature. Relaxing this assumption seems possible, but a direct adaption of the proofs would make a change of the consistency notion necessary. Let N be the number of z_i^0 that are drawn. In my asymptotic results I will let N tend to infinity. I allow that the number of possible positions K (or K_N) grows with N while I will suppress the dependency on N. This should allow actors to be more heterogeneous as the number of agents grows. Further, I assume $||M_{k1} - M_{k2}|| > \chi_N$, where $\chi_N \to 0$. So the distance between all mass points is allowed to go to zero. Hence, the mass points can also be interpreted as a grid that becomes finer with growing N. Each actor draws an M_k with probability $\mathbb{P}(z_i^0 = M_k) > c\frac{1}{K}$. Links between actors i and j are distributed according to Bernoulli distributions with probabilities

$$P_{ij}^N \equiv \theta_{z_i^0 z_j^0} \equiv \rho_N \exp(-||z_i^0 - z_j^0||).$$

Since $\exp(-||z_i - z_j||) > 0 \quad \forall z_i, z_j \in \mathcal{M}$, the expected number of outgoing edges from one node named degree is of order $\rho_N N$. Hence, ρ_N controls the degree of the graph. If ρ_N is constant, the expected degree grows proportional to N. Thinking of nodes as individuals and edges as friendships, it is unrealistic that the actual number of friends for one node grows at the same rate as individuals in the network. This would mean that the circle of friends is a significant fraction of the whole network. Therefore, assuming a ρ_N that goes to zero at a reasonable rate is appropriate in many real world applications. Many papers that are concerned with consistency of spectral clustering and related clustering methods rely on the assumption of a constant ρ_N or a ρ_N that goes to zero at a $\log(N)$-rate. Thus, working with those methods seems problematic in these settings, although clustering algorithms are computationally more efficient. A typical social network assumption is that ρ_N decreases at a rate of $\frac{1}{N^{1-\epsilon}}$. The main result will leave the choice of ρ_N open.

If ρ_N is constant, I will assume it to be smaller than one to ensure that links are not formed with probability one or higher. The definition of P_{ij}^N is based on the model of Hoff, Raftery, and Handcock 2002. It creates a model where agents who are similar in the sense of a near position in the latent space, are more likely to form links. Transitivity (if i and j and j and k are linked, then it is likely that i and k are linked) is implicit, if we model links in that manner.

The results in this chapter do not rely on the functional form of P_{ij}^N. For $P_{ij}^N = \rho_N f(||z_i^0 - z_j^0||)$ and a strictly positive differentiable function f that has a strictly positive or strictly negative derivative, the proofs work in exactly the same way. Only these attributes of the exponential function are used.

Let z^0 be the vector of true latent positions and the matrix $(A_{ij})_{N \times N}$ denotes an adjacency matrix according to z^0. This matrix contains ones and zeros that indicate which actors are linked and which are not. I will not consider self-links, therefore A_{ii} is set to zero. In addition, the edges will be undirected in the sense that the adjacency matrix is symmetric ($A_{ij} = A_{ji}$).

The Log-likelihood function writes

$$L(A,z) \equiv \sum_{i<j} A_{ij} \log(\theta_{z_i z_j}) + (1 - A_{ij}) \log(1 - \theta_{z_i z_j})$$

and the likelihood estimate $\hat{z}$ is the argmax $z \in \mathcal{M}^N$ of this function. Since the presence of a link from i to j ($A_{ij} = 1$) leads to a z_i-estimate which is closer to the z_j-estimate and $A_{jk} = 1$ to a z_k-estimate which is closer to the z_j-estimate, I automatically get a close z_i and z_k estimate. Therefore, the estimated position of a node is not only influenced by its own links, but by all connections in the network.
If I replace the A_{ij} with their expected values, I write

$$L(P,z) \equiv \sum_{i<j} P_{ij}^N \log(\theta_{z_i z_j}) + (1 - P_{ij}^N) \log(1 - \theta_{z_i z_j}).$$

This oracle version of the likelihood will be helpful to put the estimated and the true distances in relation. Furthermore, it illustrates an identification problem. As mentioned in Hoff, Raftery, and Handcock 2002, the latent positions z_i are only identified up to distance preserving transformations. This is because the above maximization only depends on the distances of the z_i and not on their true location.
Yet, I am not interested in the exact location of the z_i^0, but in their positions compared to one another. Having a good estimate of an isometry of the z_i^0 is already very informative with regard to clustering behavior of the network or centrality characteristics of a particular point. Therefore, it is worthwhile to prove a result like the next Theorem.

Theorem 1
For $b_N = (K \log(N)^{1+\zeta})^{\frac{1}{2}} |\log(\rho_N)| \rho_N^{-1} N$ and $b_N^{\frac{1}{2}}/\chi_N \le N$, an isometry T exists such that

$$\frac{1}{K\, b_N^{\frac{1}{2}}/\chi_N} \sum_{i=1}^N 1_{\{z_i^0 \neq T\, \hat{z}_i\}} = o_p(1).$$

The above result shows that the frequency of misspecified points in the ML-estimates tends to zero. This is further discussed after the next corollary. Theorem 1 is high-level in the sense that it does not specify the rates of K, χ_N and ρ_N. I consider this to be useful, since for different applications, contrasting sparsity assumptions are reasonable. For a more dense graph, a higher rate of K is possible. If one believes the nodes to have very close positions, it might be reasonable to let χ_N tend to zero faster. In the following, I will discuss three different setups, that were introduced in the literature in recent years.
The consistency concept coincides with the one of weak consistency in Zhao, Levina, and Zhu 2012. They use maximum likelihood estimation. I get the same results assuming a similar rate for ρ_N and constant K ($\rho_N = \frac{1}{N} \log(N)^2$, instead of the log they just need some $c_N \to \infty$).

Corollary 2
For $\rho_N = \frac{1}{N} \log(N)^2$ and constant K, an isometry T exists such that,

$$\frac{1}{N} \sum_{i=1}^N 1_{\{z_i^0 \neq T\, \hat{z}_i\}} = o_p(1).$$

The interpretation of this type of consistency is that the number of misspecified $\hat{z}_i$ up to an isometry does not grow as fast as the number of nodes. The ratio of misspecified agents goes to zero.
This gap becomes greater as the degree of the graph, and therefore ρ_N increases.
Choi, Wolfe, and Airoldi 2012 also use maximum likelihood estimation and allow for a growing number of blocks. If I assume the same rates for the degree of the graph and growth of the class sizes as they do, I can derive the same consistency result for K growing at a $\sqrt{\log(N)}$-rate.

Corollary 3
For $\rho_N = \frac{1}{N}\log(N)^4$ and $K = \sqrt{\log(N)}$, an isometry T exists such that,

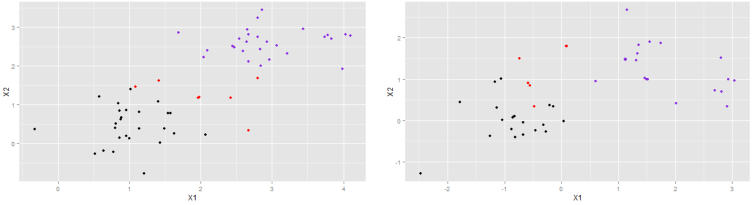
$$\frac{1}{N}\sum_{i=1}^{N} 1_{\{z_i^0 \neq T\hat{z}_i\}} = o_p(1).$$

In comparison, they derive their result for a $N^{\frac{1}{2}}$-rate for K. Yet, my result holds true for the number of misspecified points after a distance-preserving transformation, whereas their statement only detects nodes whose true class under z^0 is not in the majority within its estimated class $\hat{z}$. Therefore, they show a weak consistency result in the spirit of Zhao, Levina, and Zhu 2012 if the majority in each class is specified correctly under $\hat{z}$. An estimator up to a distance-preserving transformation is very appealing, because close points as well as central points stay close and central.
Sussman et al. 2012 use a setting which is similar to mine, in the sense that they also augment a block model with a latent space structure. The authors use a clustering method to detect class memberships. They do not assume the space to be bounded and use a clustering method to detect block affiliation. Furthermore, they leave the probability open with which a node is assigned to a mass point. The authors do not derive their main theorem for an increasing number of blocks, but explain in their possible extensions chapter how one can translate their results to this setup. Assume they use exactly my setting so a compact latent space and each mass point M_k arises with probability larger than $c\frac{1}{K}$. Assuming further like the authors do that ρ_N is constant and (to my understanding) the highest rate for K possible with their methods and these assumptions. Then, I can translate their main result to maximum likelihood estimation and conclude the following corollary.

Corollary 4
For a constant ρ_N and $K = N^{\frac{1}{7}}$, an isometry T exists such that,

$$\frac{1}{N}\sum_{i=1}^{N} 1_{\{z_i^0 \neq T\hat{z}_i\}} = o_p\Big(\frac{\log(N)}{N^{\frac{1}{4}}}\Big).$$

In this setup, Sussman et al. 2012 can only derive this expression at a $\frac{1}{N^{\frac{1}{8}}}$ -rate. Hence, I am able to prove that the number of misspecified $\hat{z}_i$ decreases faster. Nevertheless, spectral clustering like methods are computationally much more efficient. However, a major drawback of clustering methods is the assumption of a constant ρ_N which is equivalent to a linearly growing degree of the graph. Hence, one has to decide whether a restrictive setting with a constant ρ_N is reasonable in order to apply consistent clustering methods. In the next section, I will examine the performance of maximum likelihood estimation in simulations.



(a) True latent positions

(b) Estimated latent positions

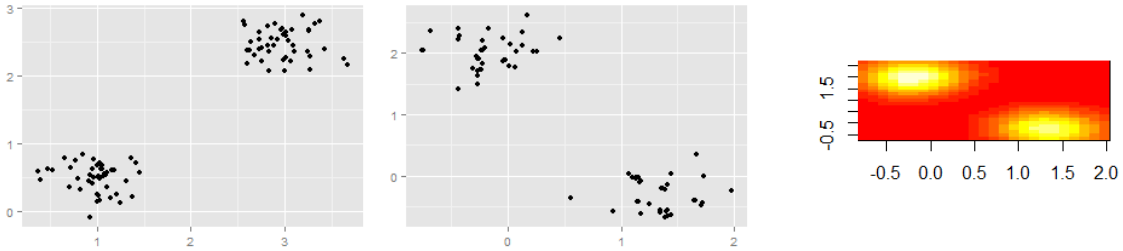
Figure 4: Figures of Latent Space True Positions and Estimates

3.3. Simulations

In this section, I want to further investigate finite sample properties of the above introduced estimation technique. Thus, I will simulate two 2-dimensional latent spaces and see how well the maximum likelihood estimator can reveal their structure. Further examples are discussed in chapter 4.

I draw 60 z_i^0 from three normal distributions $N\left(a, \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}\right)$, where a is $\binom{1}{0.5}$ or $\binom{3}{2.5}$, each with probability 0.45 and $\binom{2}{1.5}$ with probability 0.1. I fix $\rho_N = exp(-0.01)$ and use the corresponding probabilities to simulate an adjacency matrix A. The average number of links for one node is 14.8 and it ranges from 4 to 22. I can then set up the log-likelihood-function and use generalized simulated annealing to find the maximum likelihood estimate. Generalized simulated annealing is an optimization algorithm that uses jumps to detect a global optimum. These jumps are drawn from wisting distribution that is Gaussian (classical simulated annealing) or Cauchy-Lorentz (fast simulated annealing) for different parameter choices. I use the parameters recommended by Tsallis and Stariolo 1996. The corresponding true latent space has three clusters which are closely spaced. In figure 4, I indicated the different groups by colors. The points that were originally concentrated around $\binom{1}{0.5}$, $\binom{2}{1.5}$ and $\binom{3}{2.5}$ are colored black, red and blue. The simulated positions create the true latent space in figure 4(a). In this framework, we know the true dimension of the unobserved space and therefore set it to two. The estimates which are illustrated in figure 4(b) suggest that even for a small number of 60 nodes the estimates seem to resort the nodes quite well back into their clusters. The fact that the clusters of the estimated positions are correctly ordered from left to right happened by chance. As mentioned in the previous section, rotating the estimates by 180 degrees would have led to the same likelihood function value. In a further simulation, I change the diagonal elements of the variance-covariance matrix to 0.05 and pick $a = \binom{1}{0.5}$ and $a = \binom{3}{2.5}$ with probability 0.5. Thus, the true latent space has two large clusters. I draw 90 points. Here, the average number of links for a node is 32.9.

Figure 5 shows that with 90 points and a clear community structure the maximum likelihood estimator recognizes the clusters even better than in the setup with sixty points and some points between the two groups. Thus, if the number of groups is unknown, estimating latent positions can be informative.



(a) True latent positions

(b) Estimated latent positions

(c) Estimated latent positions density estimator

Figure 5: Figures of Latent Space True Positions and Estimates

3.4. Application

In this section, I illustrate the applicability of the methods introduced above using real-world data. I estimate the locations of the z_i as well as the central node between clusters for a data set that describes the hyperlinks between political blogs in the US. The illustration and discussion of the central node will follow in the next chapter. This data set was first collected by Adamic and Glance 2005 in order to measure the degree of interaction among political blogs during the 2004 U.S. election. They retrieved front pages on February 8, 2005 and February 22, 2005 and counted blog-references. Their data set contains 1494 blogs in total. In this application, I focus on the 80 most referenced blogs. The maximum number of edges for one block is 41, the minimum number is 11. On average, the number of links to each block is 29. This can be considered as a dense graph.
Adamic and Glance 2005 use self-reporting, automated categorization and manual labeling to mark their data points as liberal or conservative. Zhao, Levina, and Zhu 2012 use the same data set to illustrate ML-estimation for degree-corrected stochastic block models with a fixed number of two blocks.
They try to consistently estimate the block affiliation of each blog. Hence, they also postulate that the data set can be divided into two groups. The latent space approach improves these illustrations by giving further insights about the clustering intensity. Furthermore, it is beneficial to see which of the conservative blogs is the most liberal one and vice versa.

Following the literature on political spaces in Europe (Kriesi et al. 2006 and Bornschier 2010), I assume that two is the appropriate dimension for a latent space of political affiliation. I fix $\rho_N = exp(-0.01)$ and restrict the latent space to be distributed on $[-5,5]^2$ as in the simulations. In figure 6, I illustrate the estimates of the 80 blogs.

The estimated latent positions reinforce the classification in two clusters made by Adamic and Glance 2005 and Zhao, Levina, and Zhu 2012. I have colored the blogs that are reported to be liberal by Adamic and Glance 2005 in red. Therefore, my estimates can be understood as evidence that their labeling is also reflected in the hyperlinking behavior. As mentioned by Adamic and Glance 2005, the conservative blogs seem to have a stronger linking culture. This is reflected in the positions, because they form a cluster that is concentrated on a smaller

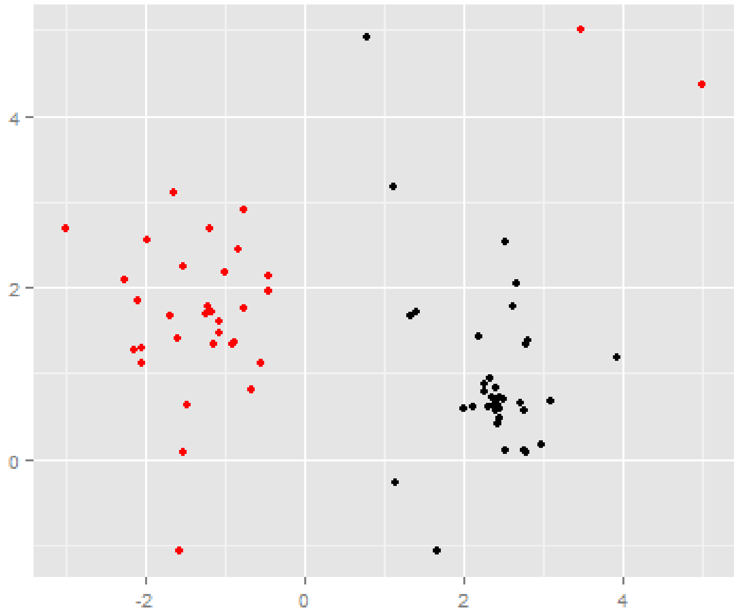


Figure 6: Top 80 Blogs Estimated Positions (2-dim)

area. Hence, the estimated probability of forming a link between two conservative blogs is higher. Since we can observe clustering into two groups, it is reasonable to look for a blog that seems to provide a balanced discussion of opinions from the different groups. This question is investigated in the next chapter.
One of the tuning parameters is the dimension of the political space. Obviously, one wants to avoid introducing unnecessary complexity. The arising question is whether a one dimensional latent space would reveal the clustering behavior equally well. Hence, I used the same method as above assuming a one dimensional space.

Inspecting the corresponding estimates in figure 7, I conclude that in the one dimensional space the grouping into conservative and liberal clusters is less obvious. Thus, if one believes the sorting of Adamic and Glance 2005 to be correct, the estimation in a one dimensional space does not seem to capture the whole picture. Furthermore, I report an illustration of estimated positions in a three dimensional latent space in the appendix. It does not substantially differ from the two dimensional case.
As a validity check, I create an adjacency matrix for the blogs with link-ranking 101 to 180. This illustration is likely to give a misleading picture, because the linking behavior to the top-80-blogs cannot be modeled here. As a consequence, no clear clustering can be identified.

3.5. Conclusion and Future Work

In this chapter, I proved a consistency result for ML-estimation in a latent space model. Furthermore, I showed how this result generalizes similar results from the block-model literature. In addition, I illustrated the usefulness of my results in an application concerned with political

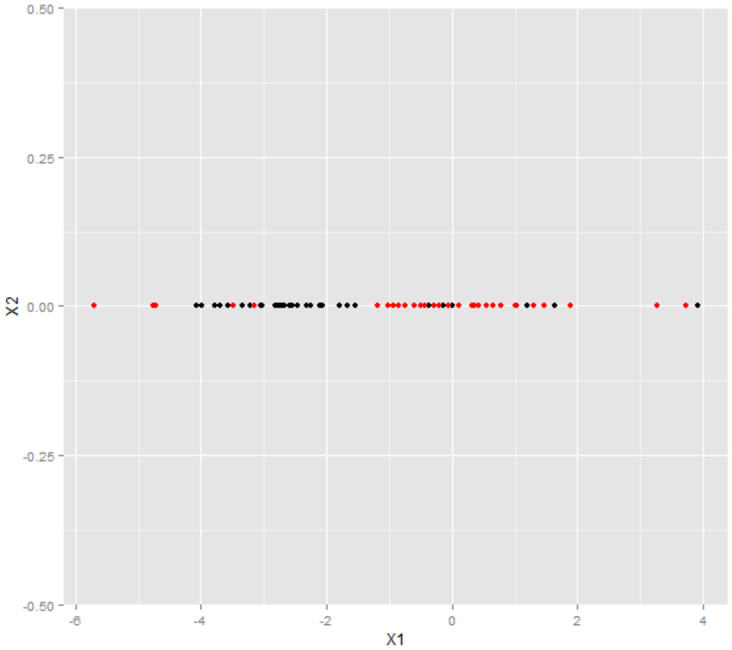


Figure 7: Top 80 Blogs Estimated Positions (1-dim)

blogs and simulations. My estimates indicate that the previous sorting into two clusters is also reflected in the blogs' linking behavior.

There are several natural extensions to the above model. Until now, I focused on undirected networks, whereas, it would be interesting to investigate how the above results translate into directed networks.

A huge weakness of maximum likelihood estimation is that its optimization is computationally very expensive. Due to those problems, it is worthwhile to look deeper into pseudo-likelihood methods (Amini et al. 2013) that improve efficiency and exploit a latent space structure as well. Deviating from Hoff, Raftery, and Handcock 2002, I do not model observed covariates explicitly which is a useful extension of the model. Nevertheless, the asymptotic results discussed above do not use covariates either.

Recently, more and more literature is concerned with so called graphon estimation (Airoldi, Costa, and Chan 2013, Wolfe and Olhede 2013, Chan and Airoldi 2014). Authors estimate the probability of a link between two nodes and assume that nodes have a characteristic on a unit interval that drives this probability. Chan and Airoldi 2014 propose a sorting algorithm to deal with identification problems.

Taking into account these non-parametric approaches, it seems reasonable to introduce a two stage density estimator and prove consistency. This density estimate could also be revealing for the overall clustering behavior.

I understand this paper as a first step to prove a consistency result in a framework where the z_i^0 are not concentrated on mass points. Unfortunately, the proofs cannot be adjusted easily. It seems possible to extend the setting to a model with weighted links by using a weighted adjacency matrix instead of the conventional one.

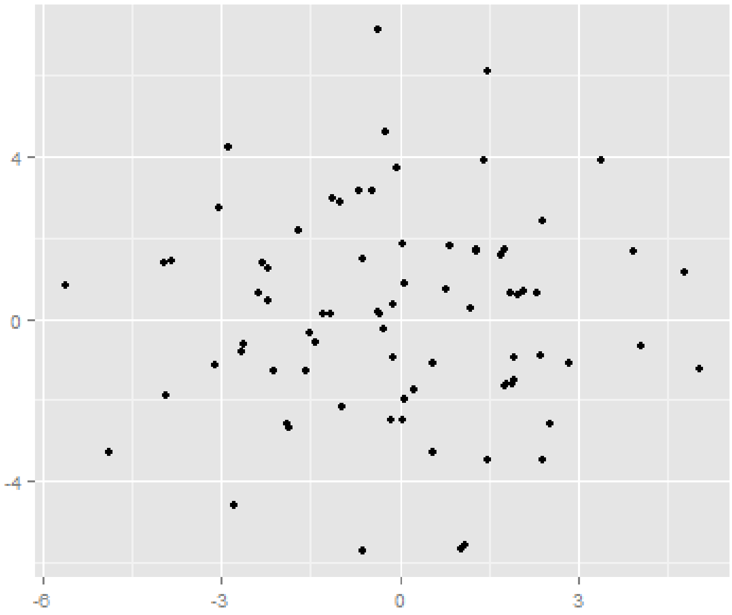


Figure 8: Blogs with Linking Ranks 101 to 180 Estimated Positions

One of my aims is to specify which nodes are in clusters and which are in between. The following chapter introduces methods to answer this question. Furthermore, I prove that the solutions are theoretically well behaved and independent of the isometry T.

4. A New Network Centrality Measure

4.1. Introduction

In this chapter, I introduce a new method to detect actors who connect clusters and therefore play an important role in a network. The method defines a novel measure of centrality that makes explicit use of a latent space structure. Furthermore, I provide theoretical evidence of asymptotical correctness of a chosen central point. In many applications it is worthwhile to identify which nodes are likely to have a well-balanced linking behavior. Examples where it is beneficial to find such an actor who has this balanced linking between groups include the revealing of good intermediaries in a market or the identification of individuals that transmit a disease. In other situations, it is worthwhile to find actors who have a neutral position and are therefore suited to resolve a conflict between two opposing groups. Ward, Siverson, and Cao 2007 stress the importance of unobserved latent positions when analyzing militarized interstate disputes. Finding a central point in these latent spaces is of interest as well. In section 4.4, I investigate internet blogs which are assumed to belong to two different political affiliations. Blogs that are located in between these camps might offer a more differentiated view.
Estimating consistently the unobserved positions with the methods of Chapter 3 allows me to distinguish which nodes share similar characteristics and can therefore be sorted into different groups. I introduce a measure of centrality that aims to identify nodes that bridge the gap between these clusters.

By augmenting a block model with a latent space, I gain structure that suggests a natural approach to reveal which nodes are inside and which nodes are between clusters. Identifying which nodes are central is not a new question in the economic networks literature. Researchers have developed many different concepts to find nodes that are central in different senses. Among many others, degree centrality, closeness, diffusion centrality (Banerjee et al. 2013), betweenness (Freeman 1977, Anthonisse 1971), Katz prestige (Katz 1953) and eigenvector centrality (Yu and Sompel 1965) are measures that are commonly used (cf. Jackson 2010). Furthermore, Bonacich 1987 and Ballester, Calvó-Armengol, and Zenou 2006 contributed algorithms to find key players. These measures declare nodes to be central depending on different notions of centrality. The authors detect nodes which lie on many shortest paths (are important for many connections that do not share a direct link) or they try to understand which nodes have many well-connected neighbors. The nature of these measures makes them likely to stay within a particular cluster, because there they are connected to many other nodes. The new measure exploits the latent space structure and intends to find nodes that are between clusters or have a balanced linking behavior between them. The latent space model allows to sort some agents into classes without forcing those nodes being in the middle to choose sides. Even if there is a clear sorting, it will remain possible to identify the actors who are closest to the other cluster. Therefore, this model is ideal to solve the above problem. In comparison to the conventional measures, the new measure uses all information available in the adjacency matrix to determine the position of a node.

In the next section, I will define the new measure of centrality on latent spaces and prove a type of consistency for this measure. In a section concerned with simulations, I compare this measure to other common centrality concepts, namely betweenness, diffusion centrality and eigenvector centrality. The new measure is tailored to identify the node that is central between distinct clusters. First simulations indicate that it is superior to the above measures in settings where this clustering structure is present. For the application discussed in Chapter 3, I detect a political blog that has a well-balanced linking behavior to blogs with different unobserved characteristics. Proofs to all results are deferred to the appendix.

4.2. Midpoints between Clusters

In this section, I introduce a concept of centrality tailored to models that have an underlying latent space structure and are estimated by maximum likelihood estimation. The measure is especially useful if agents form more than one cluster. For quite some time, many authors introduced concepts that tried to measure the centrality of nodes. Among other, they focus on concepts concerned with the number of edges starting from a node (degree-centrality) or the frequency with which nodes showed up on shortest paths (betweenness). In some applications, positions of interest are believed to cluster (Handcock, Raftery, and Tantrum 2007). The conventional measures of centrality favor points that have other points close by. The reason is that this position results in many links and more appearances on shortest paths within a particular cluster. Hence, these measures are likely to be located in a cluster. Furthermore, these measures do not necessarily use all links in the adjacency matrix. For example, the number of shortest paths for a particular actor does not necessarily change when some links are added or deleted. This might be a loss of available information. In the estimation of the latent positions, this information is implicitly used for every node.

I will use the model introduced in chapter 3. Thus, I assume that actors have positions on an unobserved Euclidean space. This offers a natural way to specify which individuals are close enough to form a cluster and which actors are halfway between these. I offer a concept that indicates which points have a well-balanced number of edges to several groups. First, one identifies clusters according to the estimated latent positions. The following definition characterizes a cluster.

Definition 2 *(γ-Cluster) A set of points forms a cluster $\mathcal{C}_\gamma$ if*

a.) for each $z_i \in \mathcal{C}_\gamma$, there exists a $z_j \in \mathcal{C}_\gamma$ such that $||z_i - z_j|| < \gamma$

b.) there exists no $z_k \notin \mathcal{C}_\gamma$ and $z_i \in \mathcal{C}_\gamma$ such that $||z_i - z_k|| < \gamma$

A reasonable extension of this definition is to bound the number of minimal members from below. This makes the definition controllable to the size of a cluster a practitioner cares about. Especially in medium sized settings, it is reasonable not to speak of a cluster when just two or three nodes are close. The minimal number of actors contained in a cluster depends on the particular application.
I derive an obvious corollary from Theorem 1 which indicates that revealing clusters $\mathcal{C}_\gamma$ of estimates is meaningful.

Corollary 5
For a γ-cluster C_γ, $b_N = (K \log(N)^{1+\zeta})^{\frac{1}{2}} |\log(\rho_N)| \rho_N^{-1} N$, $b_N^{\frac{1}{2}}/\chi_N \leq N$ and the same isometry T as in Theorem 1, it holds that

$$\frac{1}{K\, b_N^{\frac{1}{2}}/\chi_N} \sum_{z_i : \hat{z}_i \in C_\gamma} 1_{\{T^{-1} z_i^0 \in C_\gamma\}} = o_p(1).$$

Corollary 5 shows that the ratio of nodes which are not assigned to the correct cluster tends to zero. Different definitions of clusters would be possible, but lead to similar results. With this in mind, we can now calculate the midpoint of a particular cluster which pins down the location of the cluster.

Definition 3 *(Cluster Midpoint) For a cluster $\mathcal{C}_\gamma^a$ define $c^a = \frac{1}{|\mathcal{C}_\gamma^a|} \sum_{z_i \in \mathcal{C}_\gamma^a} z_i$ as the cluster midpoint of $\mathcal{C}_\gamma^a$.*

The cluster midpoints of the maximum likelihood estimate $\hat{z}$ can be related to the true z^0 in the sense of corollary 6. It shows that calculating the cluster midpoints of the $\hat{z}$ informs us about the z^0 midpoints.

Corollary 6
Let T be the isometry from Theorem 1. Then, for each cluster midpoint $\hat{c}^a$ of the $\hat{z}$-clusters C_γ^a, there exists a z^0-cluster $C_\gamma^{a,0}$ such that

$$\left\| \hat{c}^a - \frac{1}{|C_\gamma^{a,0}|} \sum_{z_i^0 \in C_\gamma^{a,0}} T^{-1} z_i^0 \right\| = o_p(1),$$

where $|C_\gamma^{a,0}|$ denotes the number of elements in $C_\gamma^{a,0}$.

This corollary illustrates that the influence of the misspecified points on the cluster midpoints does not matter asymptotically. This is true for all clusters which leads to the next definition. In the last step, I identify the node closest to the midpoint between clusters.

Definition 4 *(Midpoint between γ- Clusters) We say a node is a midpoint between clusters if its position is the closest to the arithmetic mean of cluster midpoints.*

"Closest position" means that it has the smallest Euclidean distance. Due to the concentration on mass points, the position of the midpoint between clusters is unlikely to be located exactly on the arithmetic mean between the cluster midpoints. Nevertheless, it is sensible to examine how well the new measure performs asymptotically. The following Theorem suggests that regardless of the isometry that was used in the above results, this choice is a good one.

Theorem 7
Let $\hat{\mathcal{C}}$ be the set of indices for which $\hat{z}_i = \hat{z}_{mid}$, where $\hat{z}_{mid}$ is the position of the midpoint between clusters of the estimated positions. Under the assumptions of Theorem 1 and for all $\epsilon > 0$

$$\frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\left\{ \left| \|z_i^0 - \frac{1}{A} \sum_{a=1}^A c^{a,0}\| - \|z_{mid}^0 - \frac{1}{A} \sum_{a=1}^A c^{a,0}\| \right| > \epsilon \right\}} = o_p(1),$$

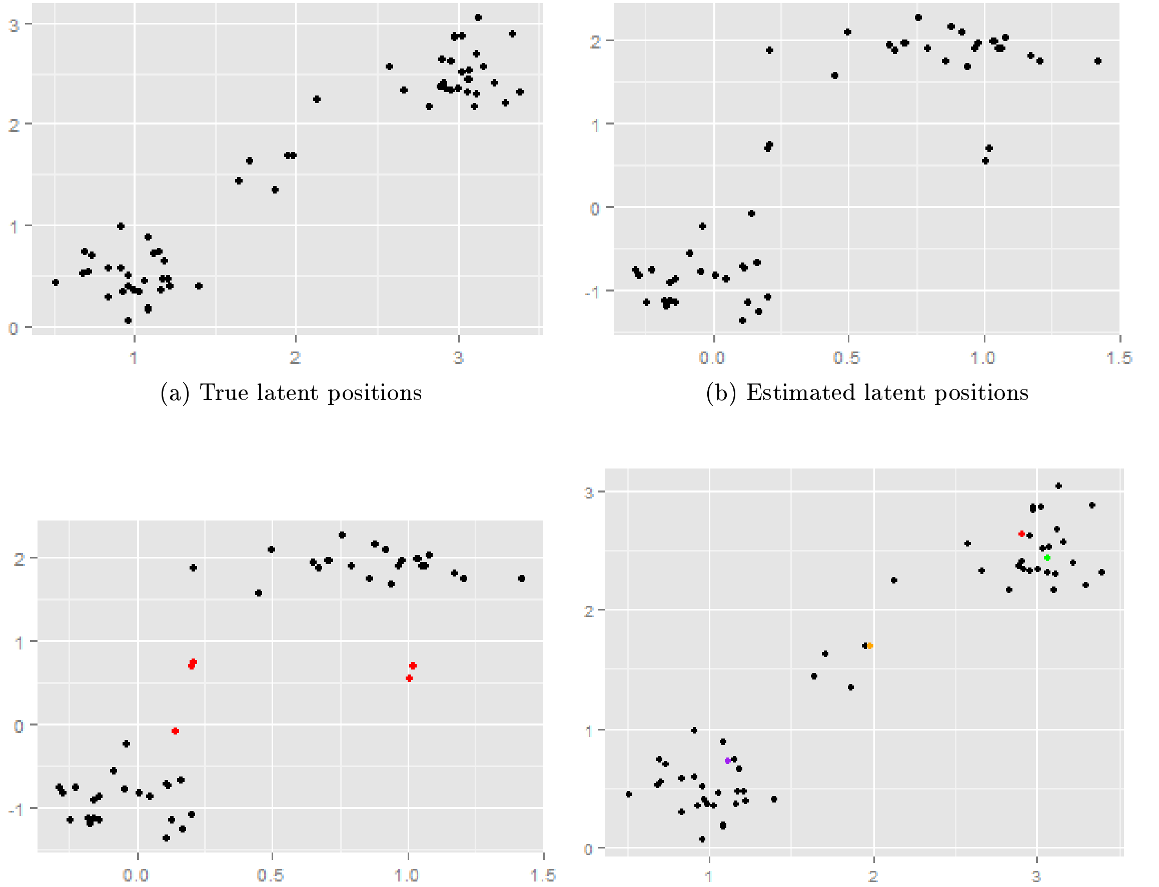
where z_{mid}^0 is the position of the midpoint between clusters of the true locations.

This statement can be understood in the following way: Take a node i that has a position associated with the midpoint between clusters. The distance of the arithmetic mean of the true clusters to the true location of i differs less than ϵ from the distance of the arithmetic mean to the true midpoint between clusters. The share of nodes for which this statement is not true tends to zero.
Thus, if the points of $\mathcal{M}$ form a grid that becomes finer, the share of nodes for which the true location of the estimated midpoint between clusters are not close to the position of the true midpoint between clusters becomes infinitely small. In other words, if one randomly picks a midpoint between clusters, the likelihood that it is far away from the true midpoint goes to zero. The statement of Theorem 7 holds true independent of the way the estimated points were transformed. Therefore, the new measure is meaningful despite the identification problem pointed out in chapter 3.

It is worthwhile to limit the number of clusters, such that one can find a node that bridges the gap between particular point clouds. One can also form the following intuition: The above procedure reweights the data according to the size of their cluster. Data points that have many other points in their neighborhood are weighted down, because they can easily be substituted. If I assume that the z_i^0 are distributed on mass points, each mass point forms a cluster. Nevertheless, mass points that are close to each other will still form clusters with more than one mass point and the above methods are reasonable.
A natural idea of a centrality measure on a latent space is to take the arithmetic mean over all $\hat{z}_i$. Again, under mild conditions one can prove that a central node determined that way is asymptotically a correct choice compared to the arithmetic mean of the true latent positions z_i^0. Nevertheless, my aim is to find a node that bridges the gap between clusters. With this goal in mind, the arithmetic mean does not work well when the number of nodes that are contained in the groups differ substantially. In this situation, the arithmetic mean would have been likely to declare a point in the bigger cluster as central. The application in this chapter can be considered as an example where it is unlikely to find equally sized clusters. As Adamic and Glance 2005 point out, the conservative blogs link more frequently. Thus, more than half of the most cited eighty blogs are conservative.
In the next chapter, I investigate how well this measure works when the underlying model is indeed a latent space model.

4.3. Simulations

In this section, I want to further examine finite sample properties of the above introduced centrality measure. It should also help to understand the difference between the new measure and other measures which are commonly used in the literature. I am motivated by scenarios where agents form two or more clusters which are connected by single actors between them. Therefore, I simulated some of these scenarios which have different location distributions of z_i^0 in a latent space. First, I assume that the latent space can be represented by $[-5,5]^2$. I draw 60 z_i^0 from three normal distributions $N\left(a, \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}\right)$, where a is $\binom{1}{0.5}$ or $\binom{3}{2.5}$, each with probability 0.45 and $\binom{2}{1.5}$ with probability 0.1. This setup results in two distinct clusters and a few points between these (see figure 9(a)). I fix $\rho_N = exp(-0.01)$ to simulate a corresponding adjacency matrix A. The probability of an edge between two nodes is usually above 0.5 if the 2 nodes are in the same cluster. If they are in two distinct ones, the probability will be around 0.1. Furthermore, the probability of a link from a point in the middle to a node in one of the clusters is between 0.15 and 0.4. The average number of links for one node is 20.6 and it ranges from 11 to 28. I can then set up the log-likelihood-function and use Generalized Simulated Annealing to find the maximum likelihood estimate. Again, I use the parameters recommended by Tsallis and Stariolo 1996. Thereafter, I identify the midpoints between γ-clusters. In this setting, I choose $\gamma = 0.7$ which is equivalent to assuming that the probability of forming a link for two nodes that connect themselves as a cluster is larger than 0.5. This is just an ad hoc criterion. A good choice of γ depends on the particular application in mind. In figure 9, I illustrate how well estimates work for this sample size.



(a) True latent positions

(b) Estimated latent positions

(c) Estimated latent positions (true mid-cluster-points (red))

(d) True latent positions with centrality points coloured

Figure 9: Figures of Latent Space True Positions and Estimates

The estimates create a picture (figure 9(b)) that reflects the overall structure with two larger clusters (figure 9(a)) quite well. Furthermore, the estimates reveal which points are located in the middle. To illustrate this, I have colored the points that were drawn from the distribution with expected value $\binom{2}{1.5}$, red (see figure 9(c)).

In 9 (d), I illustrate the performance of the different measures of centrality. As expected, betweenness, diffusion centrality and eigenvector centrality (purple, green, red) are located within one of the two clusters, whereas the midpoint between clusters (orange) coincides with a point in between. If a point in between the true groups is the point of interest, the new measure works well.
To see whether the above observations hold true under more sparse networks, I simulate two additional settings. In the first one, I choose the diagonal elements in the variance covariance matrix to be 0.3 instead of 0.05 (example also discussed in chapter 3). This creates a situation where nodes within a cluster are less likely to form links. In the second setting, I assume the midpoints of the clusters to be located at $\binom{1}{0.5}$ and $\binom{5}{4.5}$ and that the midpoints are placed around $\binom{3}{2.5}$. Hence, nodes are less likely to be connected between clusters. The average number of links for one node in the second setting decreases to 11.6 and ranges between 19 and 4. The framework that makes detection of clusters and true central points hardest is the first of the sparse settings. The true latent positions and the corresponding estimates are illustrated in figure 10 and figure 11.

Figure 10 shows that the ML-estimator still recognizes the structure of the two groups. It also illustrates the weaknesses of the new centrality measure. For my choice of $\gamma(= 0.7)$ and bounding the minimal number to three, the estimator finds three clusters. The consequential midpoint between clusters does not look convincing in the true latent space. This stresses the importance of a well chosen γ to make the concept applicable. Nevertheless, limiting the focus on two classes leads to a convincing central node. Moreover, the arithmetic mean over positions would do a good job here.
Figure 11 is in line with the results from figure 9. The two points whose true positions are located in the center of the latent space have the same estimated position. To estimate the midpoint between clusters, I focused on groups that contain three or more nodes (I detected two "clusters" with two points that were close by).

4.4. Application

In this section, I discuss how the new centrality measure can be exploited in the context of the application from chapter 3. As reviewed above the data set collected by Adamic and Glance 2005 deals with political blogs in the U.S.. The political system as well as my estimates in chapter 3 suggest that the blogs can be sorted into two clusters. In the interest of learning a differentiated point of view, it is beneficial to identify the blogs which pick up different opinions from various political blogs. The new measure of centrality seems to be tailored to exactly this problem. As above, I will use the 80 most popular blogs. The γ which defines the clusters will be set to 0.7. According to the discussion in chapter 3, the reasonable dimension of the political space seems to be two. Adamic and Glance 2005 emphasize that conservative blogs are more likely to hyperlink each others' pages. Hence, one can expect to find more conservative blogs. As mentioned above, a centrality measure based on the arithmetic mean of the estimated latent positions might be problematic to detect a neutral blog.

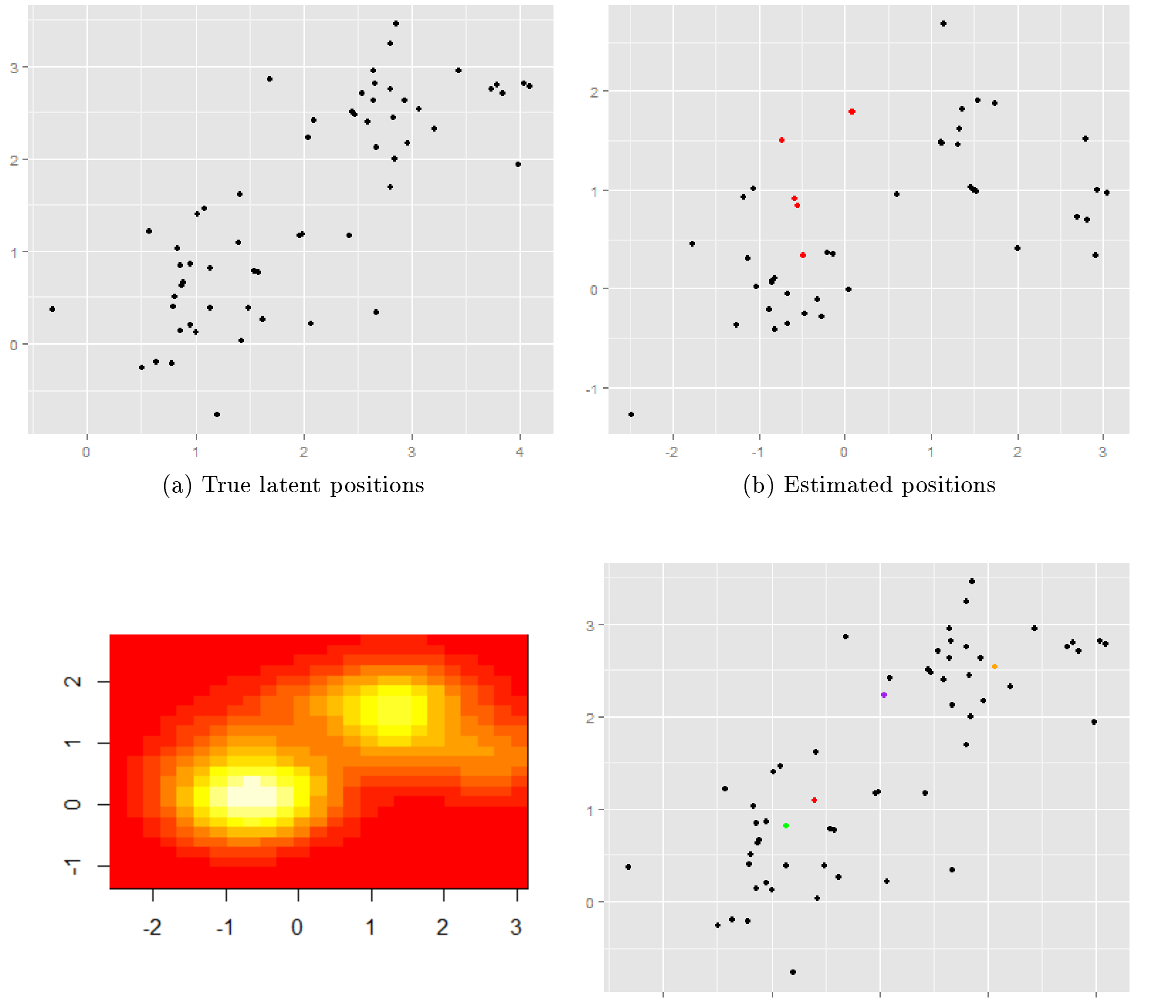
Since I do not know the true positions, I illustrate the position of the different measures of centrality in the latent space with estimated positions (see figure 12).

It is not surprising that measures which are based on the degree of the node or the neighbors are in the conservative cluster where many links are formed. According to the picture created by the estimated positions, betweeness and the midpoint between clusters seem to work well.

4.5. Conclusion and Possible Extensions

I exploited the latent space structure to introduce a new measure of centrality. I proved that it asymptotically works well and showed that it has desirable features in many finite sample applications. I used my novel measure of centrality to detect a political blog whose links are well-balanced. This blog discusses different opinions and is therefore probably more interesting than other blogs. It kind of unifies the spectrum of possible opinions, and therefore shows what is at the center of discussions.

Obviously, the measure mentioned above is not the only reasonable choice for a centrality measure which exploits the latent space structure. One option would be to use the arithmetic mean of the positions. Most of the asymptotic results will become easier to prove for this measure. I did not focus on this approach because I was also interested in bridging the gap between a huge and a small cluster.
Another possibility would be to define clusters such that no two points within a cluster are allowed to be too far away from each other. I decided against using this definition because it is likely to lead to overlapping clusters which can cause problems.
In addition, one can benefit from the estimated latent position by identifying other characteristic nodes. For example it is helpful to detect the central point of a cluster or to find the actors in a group who are farthest away from each other.



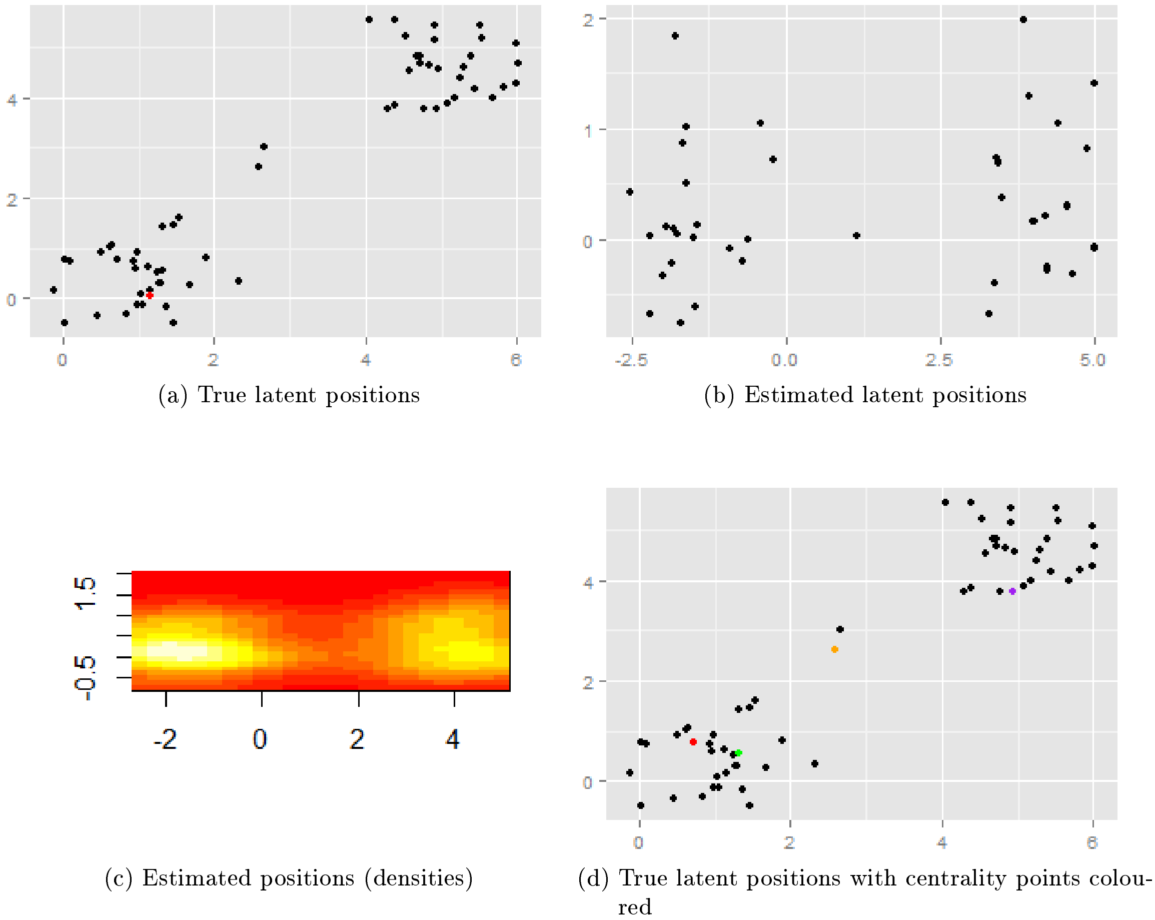
(a) True latent positions

(b) Estimated positions

(c) Estimated positions (densities)

(d) True latent positions with centrality points coloured

Figure 10: Figures of Latent Space and Estimates (Sparse Setting 1)



(a) True latent positions

(b) Estimated latent positions

(c) Estimated positions (densities)

(d) True latent positions with centrality points coloured

Figure 11: Figures of Latent Space True Positions and Estimates (Sparse Setting 2)

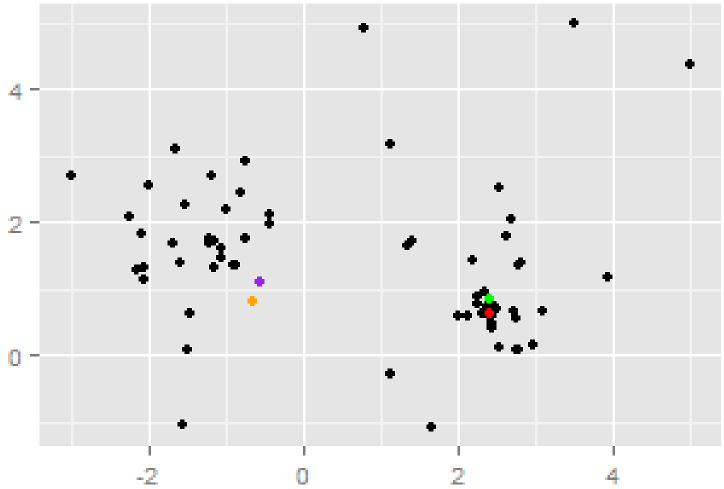


Figure 12: Top 80 Blogs Estimated Positions with Centrality Measures

Appendix

A. Definition of Estimators

Let $L_g(\cdot) = g^{-1}L(\cdot/g)$ and $K_h(\cdot) = h^{-1}K(\cdot/h)$. For the first-stage estimator set $\hat{r}_{z,j}(s) = a_0$, where a_0 satisfies

$$(a_0,\dots,a_q) \in \arg\min_{(a_0,\dots,a_q)\in\mathbb{R}^{q+1}} \sum_{i:Z_i=z,J_i=j} \big(D_i - a_0 - a_1(S_i - s) - \cdots \\ - a_q(S_i - s)^q\big)^2 L_g(S_i - s).$$

For the second-stage estimator set $\hat{m}_{z,j}(x) = b_0$, where b_0 satisfies

$$(b_0, b_1) \in \arg\min_{(b_0,b_1)\in\mathbb{R}^2} \sum_{i:Z_i=z,J_i=j} \big(Y_i - b_0 - b_1(\hat{r}_{z,j}(S_i) - x)\big)^2 K_h(\hat{r}_{z,j}(S_i) - x).$$

B. Proofs of Treatment Chapter

Proof of Proposition 2

The proposition follows from a sequence of lemmas. We first prove that the second-stage regression function and the error terms from the first- and second-stage regressions behave nicely under our assumptions about the primitives of the model.

Assumption 3 *For each $z \in \{0,1\}$*

(i) m_z is twice continuously differentiable on (x_L, x_U).

(ii) there is a positive ρ such that $\mathrm{E}_z[\exp(\rho|\zeta|) \mid S]$ and $\mathrm{E}_z[\exp(\rho|\epsilon|) \mid S]$ are bounded,

(iii) $\sigma^2_{\zeta,z}(x) = \mathrm{E}_z[\zeta^2 \mid r_z(S) = x]$, $\sigma^2_{\epsilon,z}(x) = \mathrm{E}_z[\epsilon^2 \mid r_z(S) = x]$, and $\sigma_{\epsilon\zeta,z}(x) = \mathrm{E}_z[\epsilon\zeta \mid r_z(S) = x]$ are continuous on (x_L, x_U).

Lemma 1
Assumption 1 is sufficient for Assumption 3.

Proof *The lemma follows from plugging in the structural treatment model into the observed quantities and arguing similarly to the proof of Proposition 1.*

In the next lemma we give a complete description of the relevant properties of our first-stage estimator. We provide an explicit expression of a smoothed version of the first-stage estimator that completely characterizes the impact of estimating the regressors on the asymptotic behavior of the test statistic.

Lemma 2 (First stage estimator)
The first stage local polynomial estimator can be written as

$$\hat{r}_z(s) = \rho_n(s) + R_n,$$

where

$$\sup_s |R_n| = O_p\left(g^{q+1}\sqrt{\frac{\log n}{ng}} + \frac{\log n}{ng}\right)$$

and ρ_n is given explicitly in equation (8). Wpa1 ρ_n is contained in a function class $\mathcal{R}$ that for some constant K, any $\xi > \frac{5}{4}\eta^* - \frac{1}{4}$ and all $\epsilon > 0$ can be covered by $K\exp(n^{\xi}\epsilon^{-1/2})$ ϵ-balls with respect to the sup norm. The true propensity score is contained in $\mathcal{R}$. Furthermore,

$$-m'(x)\int K_h(r_z(s) - x)(\hat{r}_z(s) - r_z(s))f_{S|Z=z}(s)\,ds = \frac{1}{n}\sum_{i:Z_i=z}\psi^{(2)}_{n,z,i}(x) + o_p(n^{-1/2}),$$

with $\psi^{(2)}_{n,z,i}$ as defined in Lemma 3. Moreover,

$$\sup_s |\hat{r}_z(s) - r_z(s)| = O_p\left(n^{-\frac{1}{2}(1-\eta^*)}\right).$$

Proof *Throughout, condition on the subsample with $Z = z$. Let $e_1 = (1, 0, \ldots, 0)^\top$ and $\mu(t) = (1, t, \ldots, t^q)^\top$. Furthermore, define*

$$\bar{M}_n(s) = \mathrm{E}\,\mu\left(\frac{S_i - s}{g}\right)\mu^\top\left(\frac{S_i - s}{g}\right)L_g(S_i - s).$$

Since we defined g in terms of the total sample size it behaves like a random variable when we work conditionally on the subsample $Z = z$. We have $g = a_n n_z^{-\eta^*} + O_p\left(n^{-\frac{1}{2}-\eta^*}\right)$ for a bounded deterministic sequence a_n. From a straightforward extension of standard arguments for the case of a deterministic bandwidth (c.f. Masry 1996) it can be shown that $\hat{r}_z$ can be written as

$$\hat{r}_z(s) = \rho_n(s) + R_n,$$

where

$$\rho_n(s) = r_z(s) + g^{q+1}b_n(s) + e_1^\top \bar{M}_n^{-1}(s)\frac{1}{n}\sum_i \mu\left(\frac{S_i - s}{g}\right)L_g(S_i - s)\zeta_i, \tag{8}$$

b_n is a bounded function and R_n has the desired order. To show that the desired entropy condition holds, note that $\bar{M}_n$ is a deterministic sequence that is bounded away from zero so that it suffices to derive an entropy bound for the functions

$$\frac{1}{n}\sum_i \mu\left(\frac{S_i - s}{g}\right)L_g(S_i - s)\zeta_i.$$

Wpa1 these functions have a second derivative that is bounded by $\sqrt{n^{-1}g^5\log n}$. The desired bound on the covering number then follows from a straightforward corollary to Theorem 2.7.1 in Vaart and Wellner 1996. To prove the statement about the smoothed first stage estimator note that under our assumptions we only have to consider the smoothed error term

$$\frac{1}{n}\sum_{i:Z_i=z}\psi^*_n(x, S_i)\zeta_i,$$

where

$$\begin{aligned}\psi_n^*(x,s) &= -\,m'(x)\int K_h(r_z(u)-x)e_1'\bar{M}_n^{-1}(u)\mu\left(\frac{s-u}{g}\right)L_g(s-u)f_{S|Z=z}(u)\,du\\ &= -\,m'(x)\int K_h(r_z(s-gu)-x)e_1'\bar{M}_n^{-1}(s-gu)\mu(u)L(u)f_{S|Z=z}(s-gu)\,du.\end{aligned}$$

Since $f_{S|Z=z}$ is bounded and has a bounded derivative there is a function $D_n(s,u)$ bounded uniformly in s, u and x satisfying

$$\bar{M}_n^{-1}(s-ug)f(s-ug)-M^{-1}=gD_n(s,u).$$

By standard kernel smoothing arguments

$$\frac{1}{n_z}\sum_{i:Z_i=z}\left\{\int K_h(r_z(S_i-ug)-x)D_n(S_i,u)\mu(u)L(u)\,du\right\}\zeta_i=O_p\left(\sqrt{\frac{\log n}{nh}}\right).$$

Noting that $L^*(u)=e_1^\top M^{-1}\mu(u)L(u)$ we have

$$\frac{1}{n}\sum_{i:Z_i=z}\psi_n^*(x,S_i)\zeta_i=\frac{1}{n}\sum_{i:Z_i=z}\psi_{n,z,i}^{(2)}(x)+o_p\big(n^{-1/2}\big).$$

Next, we give an asymptotic expansion of the integrand in (4) up to parametric order. The result states that the integrand can be characterized by a deterministic function that summarizes the deviation from index sufficiency under the alternative and an asymptotic influence function calculated under the hypothetical model $\mathcal{M}^{\text{null}}$.

Lemma 3 (Expansion)
Uniformly in x

$$\hat{m}_0(x)-\hat{m}_1(x)=\Delta_{K,h}(x)+\frac{1}{n}\sum_i\psi_{n,i}(x)+o_p\big(n^{-1/2}\big)$$

where $\psi_{n,i}=\psi_{n,i}^{(1)}+\psi_{n,i}^{(2)}$ and $\psi_{n,i}^{(j)}=\sum_{z=0,1}\psi_{n,z,i}^{(j)}$, $j=1,2$,

$$\begin{aligned}\psi_{n,z,i}^{(1)}(x) &= \frac{1_{\{Z_i=z\}}(-1)^z}{p_zf_{R,z}(x)}K_h(r_z(S_i)-x)\epsilon_i,\\ \psi_{n,z,i}^{(2)}(x) &= -\,m'(x)\frac{1_{\{Z_i=z\}}(-1)^z}{p_zf_{R,z}(x)}\int K_h(r_z(S_i-gu)-x)L^*(u)\,du\,\zeta_i.\end{aligned}$$

Here $\epsilon_i=Y^{null}-\mathrm{E}[Y^{null}\mid r_Z(S)]$, i.e, ϵ_i is the residual under the hypothetical model $\mathcal{M}^{null}$, and L^* denotes the equivalent kernel of the first step local polynomial regression.

Proof *The statement follows from an expansion of $\hat{m}_z$. Work conditionally on the subsample with $Z=z$ and let n_z denote the number of observations in the subsample. To avoid confusion, we write h_n for the second-stage bandwidth, as h will sometimes denote a generic element of a set of bandwidths. Let $h^z=n_z^{-\eta}$. Note that for C large enough h_n is contained in the set*

$$\mathcal{H}_{n_z}=\left\{h':\left|h'-h^z\right|\leq Cn_z^{-1/2-\eta}\right\}$$

wpa1. Let $e_1 = (1,0)^\top$, $\mu(t) = (1,t)^\top$ *and*

$$M_h^r(x) = \frac{1}{n}\sum_{i:Z_i=z} \mu\big((r(S_i)-x)/h\big)\mu^\top\big((r(S_i)-x)/h\big)K_h(r(S_i)-x).$$

For arbitrary $\mathbb{R}^n$*-valued random variables* W *define the local linear smoothing operator*

$$\mathcal{K}_{h,x,z}^r W = e_1^\top \left(M_h^r(x)\right)^{-1}\frac{1}{n_z}\sum_{i:Z_i=z} W_i\mu\left(\frac{r(S_i)-x}{h}\right)K_h(r(S_i)-x).$$

Decompose the estimator as

$$\begin{aligned}\hat{m}_z(x) =&\mathcal{K}_{h_n,x,z}^{\hat{r}}Y^n + \mathcal{K}_{h_n,x,z}^{\hat{r}}\left\{(Y^n - Y^{null}) - \mathrm{E}[Y^n - Y^{null} \mid S, Z]\right\}\\ &+ \mathcal{K}_{h_n,x,z}^{\hat{r}}\,\mathrm{E}[Y^n - Y^{null} \mid S, Z]\\ =&J_1 + J_2 + J_3.\end{aligned}$$

We now proceed to show that

$$\begin{aligned}&J_1 = m(x) + b_{1,n}(x) + \frac{1}{n}\sum_i\left\{\psi_{n,z,i}^{(1)}(x) + \psi_{n,z,i}^{(2)}(x)\right\} + o_p\big(n^{-1/2}\big),\\ &J_2 = o_p\big(n^{-1/2}\big),\\ &J_3 = b_{2,n}(x) + \int \mathrm{E}[\varphi_n(S,Z) \mid r_Z(S) = x + hr, Z = z]K(r)\,dr + o_p\big(n^{-1/2}\big),\end{aligned}$$

where $b_{j,n}$, $j = 1,2$, *are independent of* z *and all order symbols hold uniformly in* x*. For the* J_1 *term we apply the approach from Mammen, Rothe, and Schienle 2012 (MRS) and expand* J_1 *around the oracle estimator. Write*

$$J_1 = \mathcal{K}_{h_n,x,z}^{\hat{r}}\epsilon_i + \mathcal{K}_{h_n,x,z}^{\hat{r}}m(r_z(S_i)) = J_{1,a} + J_{1,b}.$$

For the $J_{1,a}$ *term note that* $e_1^\top\left(M_h^r(x)\right)^{-1}$ *is stochastically bounded by a uniform over* $\mathcal{H}_{n_z}$ *version of Lemma 2 in MRS. For* ρ_n *as defined in Lemma 2 write*

$$\begin{aligned}&\frac{1}{n_z}\sum_{i:Z_i=z}K_{h_n}(\hat{r}_z(S_i)-x)\epsilon_i - \frac{1}{n_z}\sum_{i:Z_i=z}K_{h_n}(r_z(S_i)-x)\epsilon_i\\ =&\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_{h_n}(\hat{r}(S_i)-x) - K_{h_n}(\rho_n(S_i)-x)\right)\epsilon_i\\ &+\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_{h_n}(\rho_n(S_i)-x) - K_{h_n}(r_z(S_i)-x)\right)\epsilon_i = I_1 + I_2.\end{aligned}$$

By the mean-value theorem $I_1 = o_p(n^{-1/2})$*. For* I_2 *note that* $\mathrm{E}_z[\epsilon \mid S] = 0$ *so that following the arguments in the proof of Lemma 2 in MRS*

$$\sup_{x;h\in\mathcal{H}_{n_z}} P\left(\sup_{r_1,r_2\in\mathcal{R}}\left|\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_h(r_1(S_i)-x) - K_h(r_2(S_i)-x)\right)\epsilon_i\right| > C^* n^{-\kappa_1}\right) \le \exp(-cn^c),$$

where κ_1 *is defined in MRS and* C^* *is a large constant. To check that* $\kappa_1 > 1/2$ *note that Theorem 1 in MRS allows bandwidth exponents in an open set so that it suffices to check the* *conditions for* h^z. *It is now straightforward to show that a polynomial number of points in* $[\underline{x}, \bar{x}] \times \mathcal{H}_{n_z}$ *provide a good enough approximation to ensure that*

$$\sup_{x,h\in\mathcal{H}_{n_z},\rho\in\mathcal{R}} \left| \frac{1}{n_z} \sum_{i:Z_i=z} \left(K_h(\rho(S_i)-x) - K_h(r_z(S_i)-x)\right)\epsilon_i \right| = O_p(n^{-\kappa_1})$$

and hence $I_2 = o_p(n^{-1/2})$. *Similar arguments apply to*

$$\frac{1}{n_z} \sum_{i:Z_i=z} \frac{\hat{r}_z(S_i)-x}{h_n} K_{h_n}(\hat{r}_z(S_i)-x)\epsilon_i.$$

Therefore, $J_{1,a}$ *can be replaced by its oracle counterpart at the expense of a remainder term that vanishes at the parametric rate:*

$$J_{1,a} = \frac{1}{n}\sum_i \psi^{(1)}_{n,z,i}(x) + o_p(n^{-1/2}).$$

Note that in the last step we also replaced n_z *by* $p_z n$. *Decompose* $J_{1,b}$ *as in the proof of Theorem 1 in MRS. It is straightforward to extend their results to hold uniformly over bandwidths in* $\mathcal{H}_{n_z}$. *Deduce that*

$$J_{1,b} = m(x) + b_{1,n}(x) - m'(x)\int K_{h_n}(r_z(s)-x)(\hat{r}_z(s)-r_z(s))f_{S|Z=z}(s)\,ds + o_p\big(n^{-1/2}\big).$$

for a sequence of functions $b_{1,n}$ *that does not depend on the design. The previous results use standard results about the Bahadur representation of the oracle estimator (cf. Masry 1996; Kong, Linton, and Xia 2010). The desired representation for* J_1 *follows from Lemma 2. For the* J_2 *term apply Lemma 2 in MRS in a similar way as described above to argue that*

$$J_2 - \mathcal{K}^{r_z}_{h_n,x,z}\left\{(Y^n - Y^{null}) - \mathrm{E}[Y^n - Y^{null} \mid S, Z]\right\} = J_2 - J_2^* = o_p\big(n^{-1/2}\big).$$

By standard kernel smoothing arguments $J_2^* = o_p\big(n^{-1/2}\big)$. *For the* J_3 *term let* $A_i = \mathrm{E}[Y_i^n - Y_i^{null} \mid S_i, Z_i]$ *and consider the behavior of the terms*

$$\frac{1}{n_z}\sum_{i:Z_i=z} A_i \left(\frac{\hat{r}_z(S_i)-x}{h_n}\right)^a K_{h_n}(\hat{r}_z(S_i)-x), \quad a = 0, 1.$$

We focus on $a = 0$. *The argument for the other case is similar. Let* $K'_h(\cdot) = h^{-1}K'(\cdot/h)$. *For any* $\tilde{r}$ *(pointwise) between* $\hat{r}_z$ *and* r_z

$$\sup_x \left| \frac{1}{n_z}\sum_{i:Z_i=z} K'_{h_n}(\tilde{r}(S_i)-x) \right| \le C \sup_x \frac{1}{n_z h^z} \sum_{i:Z_i=z} 1_{\{|r_z(S_i)-x| \le Ch^z\}} = O_p(1)$$

for a positive constant C. *Noting that* $\max_{i\le n}|A_i| = O_p(c_n)$ *it is now easy to see that*

$$\begin{aligned}
&\frac{1}{n_z}\sum_{i:Z_i=z} A_i K_{h_n}(\hat{r}_z(S_i)-x) \\
=&\frac{1}{n_z}\sum_{i:Z_i=z} A_i \left[K_{h_n}(r_z(S_i)-x) + K'_{h_n}(\tilde{r}(S_i)-x)\frac{\hat{r}_z(S_i)-r_z(S_i)}{h_n}\right] \\
=&\frac{1}{n_z}\sum_{i:Z_i=z} A_i K_{h_n}(r_z(S_i)-x) + o_p\big(n^{-1/2}\big)
\end{aligned}$$

uniformly in x. Let $M = \int \mu(t)\mu^\intercal(t)K(t)\,dt$, $M_n = M_{h_n}^{r_z}$ and $\bar{M}_n = \mathrm{E}\,M_n$. By Lemma 2 in Mammen, Rothe, and Schienle 2012 and standard arguments we have

$$\begin{aligned}M_{h_n}^{\hat{r}_z}(x) - f_{R|Z=z}M =& M_{h_n}^{\hat{r}_z}(x) - M_n(x) + M_n(x) - \bar{M}_n(x) + \bar{M}_n(x) - f_{R|Z=z}(x)M \\ =& O_p\left(n^{-\frac{1}{2}(1-3\eta)} + \sqrt{\frac{\log n}{nh_n}} + h_n\right)\end{aligned}$$

uniformly in x. Therefore,

$$J_3 - f_{R|Z=z}^{-1}(x)\frac{1}{n_z}\sum_{i:Z_i=z} A_i K_{h_n}(r_z(S_i) - x) = J_3 + J_3^* = o_p\big(n^{-1/2}\big).$$

It is straightforward to show that under our assumptions J_3^* can be replaced by its expectation at the expense of an uniform $o_p(n^{-1/2})$ term. Since

$$\mathrm{E}[Y^n - Y^{null} \mid S, Z] = \mathrm{E}[Y^n - Y^{null} \mid r_Z(S)] + \varphi_n(S, Z),$$

and since $f_{R|Z=z}$ has a bounded derivative

$$\begin{aligned}\mathrm{E}_z\, J_3^* =& \int \mathrm{E}[Y^n - Y^{null} \mid r_Z(S) = x + h_n r]K(r)\,dr \\ &+ \int \mathrm{E}[\varphi_n(S,Z) \mid r_Z(S) = x + h_n r, Z = z]K(r)\,dr + o\big(n^{-1/2}\big).\end{aligned}$$

Here we keep implicit that we are treating h_n as a constant in the above expectations, i.e., we are integrating with respect to the marginal measure of (Z, S). The conclusion follows by noting that the first term on the right-hand side is independent of z.

Plugging in from Lemma 3 gives an asymptotic expansion of the test statistic.

Lemma 4

$$T_n = T_{n,a} + T_{n,b} + \int \Delta_{K,h}^2(x)\,dx + o_p(n\sqrt{h}),$$

where

$$T_{n,a} = \frac{2}{n^2}\sum_{i<j}\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx \quad \textit{and} \quad T_{n,b} = \frac{1}{n^2}\int \sum_i \psi_{n,i}^2(x)\,dx.$$

Proof *Plug in from Lemma 3, expand the square and inspect each term separately.*

Lemma 5 (Variance)
For $T_{n,a}$ as defined in Lemma 4

$$\begin{aligned}&\mathrm{var}(T_{n,a}) = n^{-2}h^{-1}V + o\left(n^{-2}h^{-1}\right) \quad \textit{and} \\ &n\sqrt{h}T_{n,a} \xrightarrow{d} \mathcal{N}(0, V).\end{aligned}$$

Proof *For the first part of the lemma, note that*

$$\int K_h(r_z(s-gu)-x)L^*(u)\,du = \int \{K_h(r_z(s)-x) + \underbrace{K'(\chi_1/h)\partial_s r_z(\chi_2)u}_{\equiv a(s,u,x)}\frac{g}{h^2}\}L^*(u)\,du,$$

where χ_1 *is an intermediate value between* $r_z(s-hu)-x$ *and* $r_z(s)-x$*, and* χ_2 *is an intermediate value between* $s-hu$ *and* s*. As* K *and* r_z *have bounded derivatives*

$$\tilde{a}(r,x) = \mathrm{E}\Big[\int a(S,u,x)L^*(u)\,du \mid r_z(S) = r\Big]$$

is a bounded function. By standard U-statistic arguments

$$\begin{aligned}\mathrm{var}\left(2\sum_{i<j}\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx\right) &= 4\sum_{i<j}\mathrm{E}\Big[\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx\Big]^2\\ &= 4\binom{n}{2}\int h\left\{\mathrm{E}[\psi_{n,1}(x)\psi_{n,1}(x+hx')]\right\}^2\,dx'\,dx.\end{aligned}$$

Note that

$$\mathrm{E}[\psi_{n,1}(x)\psi_{n,1}(x+hx')] = \sum_{z\in\{0,1\}}\mathrm{E}[\psi_{n,z,1}(x)\psi_{n,z,1}(x+hx')].$$

We consider here only one of the terms composing $\mathrm{E}[\psi_{n,z,1}(x)\psi_{n,z,1}(x+hx')]$*. For the other terms similar arguments apply. Let*

$$q(x) = -\frac{m'(x)1_{\{Z=z\}}}{p_z f_{R|Z=z}}\int K_h(r_z(S-gu)-x)L^*(u)\,du.$$

Using $\mathrm{E}_z[\zeta^2 \mid r_Z(S) = x] = x(1-x)$ *we have*

$$\begin{aligned}&h[\mathrm{E}\,q(x)q(x+hx')\zeta_1^2]\\ =&h\,\mathrm{E}\Bigg[\frac{1_{\{Z=z\}}m_z'(x)m_z'(x+hx')}{p_z^2 f_{R|Z=z}(x)f_{R|Z=z}(x+hx')}(K_h(r_z(S)-x)+\frac{g}{h^2}\tilde{a}(r_z(S),x))\cdots\\ &\cdots(K_h(r_z(S)-x-hx')+\frac{g}{h^2}\tilde{a}(r_z(S),x-hx')\zeta^2\Bigg]\\ =&\frac{x(1-x)[m_z'(x)]^2}{p_z f_{R|Z=z}(x)}\int K(y)K(x'-y)\,dy + o(1) = \frac{x(1-x)[m_z'(x)]^2}{p_z f_{R|Z=z}(x)}K^{(2)}(x') + o(1).\end{aligned}$$

For the second part of the lemma it suffices to check the two conditions of Theorem 2.1 in de Jong 1987. Let $W_{ij} = 2n^{-1}\sqrt{h}\int \psi_i(x)\psi_j(x)$ *and show that*

$$\begin{aligned}\mathrm{var}^{-1}\Big(\sum_{i<j}W_{ij}\Big)\max_{1\le i\le n}\sum_{1\le j\le n}\mathrm{var}(W_{ij}) &\to 0\\ \mathrm{var}^{-2}\Big(\sum_{i<j}W_{ij}\Big)\mathrm{E}\Big\{\sum_{i<j}W_{ij}\Big\}^4 &\to 3.\end{aligned}$$

The first condition holds trivially. To show that the second condition is satisfied note that $\mathrm{var}(\sum_{i<j} W_{ij})$ *converges to a constant. It is easy to see that asymptotically only terms of the form* $\mathrm{E}\, W_{ij}^2 W_{kl}^2$ *with* $\{i,j\} \cap \{k,l\} = \varnothing$ *will contribute to* $\mathrm{E}\big[\sum_{i<j} W_{ij}\big]^4$*. There are*

$$\binom{4}{2} \frac{\binom{n}{2}\left[\binom{n}{2} - 1\right]}{2!} \approx \frac{3}{4} n^4$$

such terms when expanding $\mathrm{E}\big[\sum_{i<j} W_{ij}\big]^4$*. The condition then follows by noting that*

$$\mathrm{var}\Big(\sum_{i<j} W_{ij}\Big) = \sum_{i<j} \mathrm{E}\, W_{ij}^2$$

and that $\mathrm{E}\, W_{ij}^2 W_{kl}^2$ *factors as* $\mathrm{E}\, W_{ij}^2\, \mathrm{E}\, W_{kl}^2$*.*

We now apply standard U-statistic theory. As the next two lemmas show, $T_{n,b}$ contributes to the asymptotic bias and $T_{n,b}$ contributes to the asymptotic variance.

Lemma 6 (Bias)
For $T_{n,b}$ *as defined in Lemma 4*

$$n\sqrt{h} T_{n,b} = \frac{1}{\sqrt{h}} \gamma_n + o_p(1),$$

where γ_n *is a deterministic sequence converging to* γ*.*

Proof *Write*

$$n\sqrt{h} T_{n,b} = \frac{\sqrt{h}}{n} \sum_i \int \psi_{n,i}^2(x)\, dx = \mathrm{E}\left\{\frac{\sqrt{h}}{n} \sum_i \int \psi_{n,i}^2(x)\, dx\right\} + o_p(1) \equiv \gamma_n + o_p(1).$$

Define the function a as in the proof for Lemma 5. To compute γ_n *write*

$$\begin{aligned}
\psi_{n,z,i}^2(x) =& \frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)} \Big\{ K_h^2(r_z(S) - x)\epsilon^2 + [m'(x)]^2 K_h^2(r_z(S) - x)\zeta^2 \\
& - 2m'(x) K_h(r_z(S) - x)\epsilon\zeta \Big\} \\
& + \frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)} \left(\frac{g}{h^2} \int g(S,u,x) L^*(u)\, du\right)^2 \zeta^2 + \\
& \frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)} \frac{g}{h^2} \left(\int g(S,u,x) L^*(u)\, du\right) K_h(r_z(S) - x)\epsilon\zeta \\
=& \Gamma_1(S,x) + \Gamma_2(S,x) + \Gamma_3(S,x).
\end{aligned}$$

Note that

$$h \sum_{z=0,1} \mathrm{E} \int \Gamma_1(S,x)\, dx \to \gamma,$$

where we kept the dependence of Γ_1 *on z implicit. Now show that the other terms entering* γ_n *vanish. To show that* $h \sum_{z=0,1} \mathrm{E} \int \Gamma_3(S,x)\, dx \to 0$ *it suffices to show that*

$$\mathrm{E}_z\left[\left(\int g(S,u,x) L^*(u)\, du\right) \epsilon\zeta \mid r_z(S)\right]$$

is bounded. This follows immediately from the fact that $\int g(S,u,x)L^*(u)\,du$ *is bounded and hence*

$$\mathrm{E}_z\left[\int g(S,u,x)L^*(u)\,du\,\epsilon\zeta \mid r_z(S)\right] \lesssim \mathrm{E}_z[|\epsilon\zeta| \mid r_z(S)] \le \sqrt{\sigma_\epsilon^2(r_z(S))} \le C$$

for some constant C. *For* $h\sum_{z=0,1}\mathrm{E}\int\Gamma_2(S,x)\,dx$ *argue similarly.*

Proof of Proposition 3

Proof *Using the expansion from Lemma 3 and applying standard smoothing arguments to the stochastic term we get that for a small enough open set* $\mathcal{G}_x \supset [\underline{x},\bar{x}]$

$$\sup_{x\in\mathcal{G}_x}|\hat{m}_0(x)-\hat{m}_1(x)|^2 = O\left(\frac{1}{n\sqrt{h}}+g^{2(q+1)}\right)+O_p\left(\frac{\log n}{nh}\right)+o_p\left(\frac{1}{n}\right).$$

Write

$$T_n(\underline{x}_n,\bar{x}_n)-T_n(\underline{x},\bar{x}) = T_n(\underline{x}_n,\underline{x})-T_n(\bar{x},\bar{x}_n).$$

We can bound $T_n(\underline{x}_n,\underline{x})$ *by*

$$|\underline{x}_n-\underline{x}|\sup_{x\in\mathcal{G}_x}|\hat{m}_0(x)-\hat{m}_1(x)| = o_p(n\sqrt{h}).$$

Similarly, we can find a bound for $T_n(\bar{x},\bar{x}_n)$.

C. Proofs of Network Chapters

Proof (Proof of Theorem 1) *I start by using similar lemmas and proofs as in Choi, Wolfe, and Airoldi 2012.*

Lemma 8
For $a_N = (K\,\log(N)^{1+\zeta}\,\rho_N\,M)^{\frac{1}{2}}|\log(\rho_N)|$ *and* $\zeta > 0$,

$$\max_z|L(A,z)-L(P,z)| = o_P(a_N).$$

Proof *Under the above assumptions I conclude,*

$$L(A,z)-L(P,z) = X-\mathbb{E}[X]$$

where $X = \sum_{i<j}A_{ij}\,\log\left(\frac{\theta_{z_iz_j}}{1-\theta_{z_iz_j}}\right)$.

$$\begin{aligned}L(A,z)-L(P,z) &= \sum_{i<j}(A_{ij}-P_{ij}^N)\log(\theta_{z_iz_j})-(A_{ij}-P_{ij}^N)\log(1-\theta_{z_iz_j})\\ &= \sum_{i<j}A_{ij}\,\log\left(\frac{\theta_{z_iz_j}}{1-\theta_{z_iz_j}}\right)-\sum_{i<j}P_{ij}^N\,\log\left(\frac{\theta_{z_iz_j}}{1-\theta_{z_iz_j}}\right)\\ &= X-\mathbb{E}[X].\end{aligned}$$

Since $C_1\rho_N \le P_{ij}^N \le C_2\,\rho_N$ *and therefore* $C_1\rho_N \le \theta_{z_iz_j} \le C_2\rho_N$, $A_{ij}\,\log\left(\frac{\theta_{z_iz_j}}{1-\theta_{z_iz_j}}\right)$ *is bounded by a factor* $C\log(\rho_N)$. *Using a Bernstein inequality due to Chung and Lu 2006, I deduce*

$$\mathbb{P}(|X-\mathbb{E}[X]| \ge \epsilon) \le 2\,\exp\left(-\frac{\epsilon^2}{2\sum_{i<j}\mathbb{E}[X_{ij}^2] + \frac{2}{3}\,\epsilon\,C}\right).$$

Since $\sum_{i<j}\mathbb{E}[X_{ij}^2] \le \sum_{i<j}(P_{ij}^N)^2\,C_0\log(\rho_N)^2 \le C_0\log(\rho_N)^2\,\rho_N^2\sum_{i<j}\exp(-||z_i-z_j||) \simeq \rho_N^2\,N^2\,\log(\rho_N)^2$,

$$\begin{aligned}\mathbb{P}(\max_z|X-\mathbb{E}[X]| \ge a_N\,\epsilon) &\le N^K\,2\,\exp\left(-\frac{a_N^2\epsilon^2}{2\sum_{i<j}\mathbb{E}[X_{ij}^2] + \frac{2}{3}\,\epsilon\,a_N\,C}\right)\\ &\le 2\,\exp\left(K\,\log(N) - \frac{a_N^2\epsilon^2}{2\sum_{i<j}\mathbb{E}[X_{ij}^2] + \frac{2}{3}\,\epsilon\,a_N\,C}\right)\\ &\to 0.\end{aligned}$$

Lemma 9
For $a_N = (K\,\log(N)^{1+\zeta}\,\rho_N\,M)^{\frac{1}{2}}|\log(\rho_N)|$ *and* $\zeta > 0$,

$$L(P,z^0) - L(P,\hat{z}) = o_P(a_N).$$

Proof $\hat{z}$ *is the maximum of the log-likelihood function* $L(A;\hat{z}) \ge L(A;z^0)$, *this implies* $L(A,\hat{z}) = L(A,z^0) + \delta$ *for* $\delta \ge 0$. *I deduce*

$$\begin{aligned}L(P,z^0) - L(P,\hat{z}) &\le |L(P,z^0) - L(P,\hat{z}) + \delta|\\ &\le |L(P,z^0) - L(A,z^0)| + |L(A,z^0) + \delta - L(P,\hat{z})|\\ &= |L(P,z^0) - L(A,z^0)| + |L(A,\hat{z}) - L(P,\hat{z})|\\ &= o_p(a_N).\end{aligned}$$

Lemma 10
For $b_N = \frac{a_N}{\rho_N^2} = (K\,\log(N)^{1+\zeta})^{\frac{1}{2}}|\log(\rho_N)|\,\rho_N^{-1}\,N$

$$\sum_{i<j}(||\hat{z}_i - \hat{z}_j|| - ||z_i^0 - z_j^0||)^2 = o_p(b_N)$$

Proof *I know from Lemma 9 that* $L(P,z^0) - L(P,\hat{z}) = o_P(a_N)$. *In addition, one can derive*

$$\begin{aligned}L(P,z^0) - L(P,\hat{z}) &= \sum_{i<j} KL(\theta_{z_i^0z_j^0}||\theta_{\hat{z}_i\hat{z}_i})\\ &\ge \sum_{i<j} 2\,(\theta_{z_i^0z_j^0} - \theta_{\hat{z}_i\hat{z}_i})^2\\ &= \sum_{i<j}\,2\,\rho_N^2\,(\exp(-||z_i^0 - z_j^0||) - \exp(-||\hat{z}_i - \hat{z}_j||))^2\\ &= \sum_{i<j}\,2\,\rho_N^2\,\exp(2\,\tilde{z}_{ij})(||\hat{z}_i - \hat{z}_j|| - ||z_i^0 - z_j^0||)^2,\end{aligned}$$

where $\tilde{z}_{ij}$ *is a value between* $-||\hat{z}_i - \hat{z}_j||$ *and* $-||z_i^0 - z_j^0||$. *Since the latent space is bounded and* $2\,\exp(2\,\tilde{z}_{ij}) \ge 2\,\exp(-2\,\mathit{diam}(\mathit{Latent\ Space})) \ge C$, *I deduce* $\sum_{i<j}(||\hat{z}_i - \hat{z}_j|| - ||z_i^0 - z_j^0||)^2 = o_p(b_N)$

All points on the latent space arise with probability larger than $c\frac{1}{K}$. Therefore with probability tending to one, for each point M_k ($k = 1, \dots, K$) there exists a subsequence $(z_n^0)_{n \in I_k}$ (here I regard $z^0 = z_n^0$ as sequence $(z_n^0)_{n \in \{1,\dots,N\}}$) such that $z_n^0 = M_k$ for all $n \in I_k$ and elements in I_k do not grow faster than $(b_n^{\frac{1}{2}}/(K\chi_n))_{n \in \{1,\dots,N\}}$ (let $i_l \in I_k \quad \forall l$ and $i_1 < i_2 < \dots$, then $\limsup_{l \to \infty} \frac{i_l}{b_l^{\frac{1}{2}}/(K\chi_l)} \leq C$). Speaking more loosely the frequency with which M_k occurs in $(z_n^0)_{n \in \{1,\dots,N\}}$ is higher than $b_n^{\frac{1}{2}}/(K\chi_n)$.

Since there are K possible realizations of one $\hat{z}_i$, there needs to be one point that occurs with frequency $b_n^{\frac{1}{2}}/\chi_n$ or higher in the subsequence $(\hat{z}_n)_{n \in I_k}$, where I_k is the same index set as above. Furthermore, I will show there is at most one point.

Assume $(\hat{z}_n)_{n \in I_k}$ has two points, z_1 and z_2, that are realized on index sets $I_{k_1} \subset I_k$ and $I_{k_2} \subset I_k$ with frequency $b_n^{\frac{1}{2}}/\chi_n$ such that $\|z_1 - z_2\| > \chi_n$. But then

$$\begin{aligned}
\sum_{i<j} (\|\hat{z}_i - \hat{z}_j\| - \|z_i^0 - z_j^0\|)^2 &\geq \sum_{i \in I_{k_1}, j \in I_{k_2}} (\|\hat{z}_i - \hat{z}_j\| - \|z_i^0 - z_j^0\|)^2 \\
&= \sum_{i \in I_{k_1}, j \in I_{k_2}} \chi_N^2 \\
&\gtrsim b_N,
\end{aligned}$$

where $\gtrsim$ denotes that the term on the left goes to infinity faster than the term on the right. Hence, there is only one z_k that occurs with frequency $b_n^{\frac{1}{2}}/\chi_N$ or higher. This procedure can be redone for all k. Set $T z_k \equiv M_k$ for all k. If T is defined in that manner, it is bijective.

It remains to show T is an isometry and that $T\hat{z}$ fulfills the consistency concept.

T is an Isometry:

Assume there would be z_1 and z_2 in $\mathcal{M}$ such that

$$\|z_1 - z_2\| \neq \|T z_1 - T z_2\|.$$

I know all points M_k have a unique T-preimage z_k, hence, there exist sequences $(\hat{z}_n)_{n \in I_1} = z_1$ and $(\hat{z}_n)_{n \in I_2} = z_2$ each with frequency $b_n^{\frac{1}{2}}$ or higher. Thus, I know

$$\|\hat{z}_{n1} - \hat{z}_{n2}\| - \|T\hat{z}_{n1} - T\hat{z}_{n2}\| > \beta,$$

for $\beta > 0$. For each subsequence $(\hat{z}_n)_{n \in I_1}$ there exists M_{k1} such that $(z_n^0)_{n \in I_1^s}$ is equal to M_{k1} for each element and fulfills the same frequency restrictions. This then needs to be the image *of z_k. I deduce*

$$\begin{aligned}
\sum_{i<j}(\|\hat{z}_i - \hat{z}_j\| - \|z_i^0 - z_j^0\|)^2 &\geq \sum_{i\in I_1^s, j\in I_2^s} (\|\hat{z}_i - \hat{z}_j\| - \|z_i^0 - z_j^0\|)^2 \\
&= \sum_{i\in I_1^s, j\in I_2^s} (\|\hat{z}_i - \hat{z}_j\| - \|M_{k1} - M_{k2}\|)^2 \\
&= \sum_{i\in I_1^s, j\in I_2^s} (\|\hat{z}_i - \hat{z}_j\| - \|T\,\hat{z}_i - T\,\hat{z}_j\|)^2 \\
&= \sum_{i\in I_1^s, j\in I_2^s} (\|z_1 - z_2\| - \|T\,z_1 - T\,z_2\|)^2 \\
&\geq \sum_{i\in I_1^s, j\in I_2^s} \beta^2 \\
&\gtrsim b_N.
\end{aligned}$$

T fulfills the consistency concept:
For the above T , it holds that

$$\begin{aligned}
\frac{1}{K\,b_N^{\frac{1}{2}}/\chi_N}\sum_{i=1}^{N} 1_{\{z_i^0 \neq T\,\hat{z}_i\}} &= \frac{1}{K\,b_N^{\frac{1}{2}}/\chi_N}\sum_{k=1}^{K}\sum_{z_i^0 : z_i^0 = M_k} 1_{\{z_i^0 \neq T\,\hat{z}_i\}} \\
&= o_P(1).
\end{aligned}$$

Proof (Proof of Corollary 6) *For a cluster C_γ^a, let $\mathcal{Z}_a$ be the subset that contains all z_k, where there exists an a $\hat{z}_i$ such that $\hat{z}_i \in C_\gamma^a$ and $\hat{z}_i = z_k$. To the set $\mathcal{Z}_a$, there exists a T-preimage $\mathcal{M}_{\mathcal{Z}_a} = \{M_k : z_k \in \mathcal{Z}, T^{-1}z_k = M_k\}$.*
I can use the same arguments as in the proof of theorem 1 to show that $\mathcal{M}_{\mathcal{Z}_a}$ also form a γ-cluster $C_\gamma^{a,0}$.
For $z_i^0 \in \mathcal{M}_{\mathcal{Z}_a}$, I can deduce

$$\begin{aligned}
\left\|\hat{c}^a - \frac{1}{|C_\gamma^{a,0}|}\sum_{z_i^0\in C_\gamma^{a,0}} T^{-1}z_i^0\right\| &= \left\|\frac{1}{|C_\gamma^a|}\sum_{\hat{z}_i\in C_\gamma^a}\hat{z}_i - \frac{1}{|C_\gamma^{a,0}|}\sum_{z_i^0\in C_\gamma^{a,0}} T^{-1}z_i^0\right\| \\
&\leq \underbrace{\left\|\frac{1}{|C_\gamma^a|}\sum_{\hat{z}_i\in C_\gamma^a}\hat{z}_i - \frac{1}{|C_\gamma^{a,0}|}\sum_{\hat{z}_i\in C_\gamma^a}\hat{z}_i\right\|}_{(1)} \\
&\quad + \underbrace{\left\|\frac{1}{|C_\gamma^{a,0}|}\sum_{\hat{z}_i\in C_\gamma^a}\hat{z}_i - \frac{1}{|C_\gamma^{a,0}|}\sum_{z_i^0\in C_\gamma^{a,0}} T^{-1}z_i^0\right\|}_{(2)}.
\end{aligned}$$

Since all K points occur with the same frequency as seen in the proof of theorem 1, $\frac{1}{|C_\gamma^a|}$ and $\frac{1}{|C_\gamma^{a,0}|}$ tend to zero with the same rate. Therefore, (1) is an $o_p(1)$-term.

For (2), I can derive

$$
\begin{aligned}
&\left\| \frac{1}{|C_\gamma^{a,0}|} \sum_{\hat{z}_i \in C_\gamma^a} \hat{z}_i - \frac{1}{|C_\gamma^{a,0}|} \sum_{z_i^0 \in C_\gamma^{a,0}} T^{-1} z_i^0 \right\| \\
&\qquad = \left\| \frac{1}{|C_\gamma^{a,0}|} \sum_{\hat{z}_i \in C_\gamma^a} \hat{z}_i \, 1_{\{z_i^0 \neq T\,\hat{z}_i\}} - \frac{1}{|C_\gamma^{a,0}|} \sum_{z_i^0 \in C_\gamma^{a,0}} T^{-1} z_i^0 \, 1_{\{z_i^0 \neq T\,\hat{z}_i\}} \right\| \\
&\qquad \leq \left\| \frac{1}{|C_\gamma^{a,0}|} \sum_{\hat{z}_i \in C_\gamma^a} C \, 1_{\{z_i^0 \neq T\,\hat{z}_i\}} \right\| + \left\| \frac{1}{|C_\gamma^{a,0}|} \sum_{z_i^0 \in C_\gamma^{a,0}} C \, 1_{\{z_i^0 \neq T\,\hat{z}_i\}} \right\| \\
&\qquad = o_p(1),
\end{aligned}
$$

where the last step immediately follows from Theorem 1.

Proof (Proof of Theorem 7) *Since z_{mid}^0 is by definition minimizer over all points of $\mathcal{M}$ of $||z_{mid}^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}||$, I can rewrite the claimed statement as*

$$\frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\{||z_i^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| - ||z_{mid}^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| > \epsilon\}} = o_P(1).$$

Let T be the isometry from Theorem 1, I can write

$$
\begin{aligned}
&\frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\{||z_i^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| - ||z_{mid}^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| > \epsilon\}} \\
&\quad \leq \frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\{||z_i^0 - T\,\hat{z}_{mid}|| + ||T\,\hat{z}_{mid} - T[\frac{1}{A}\sum_{a=1}^A \hat{c}^a]|| + ||T[\frac{1}{A}\sum_{a=1}^A \hat{c}^a] - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| - ||z_{mid}^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| > \epsilon\}} \\
&\quad \leq \underbrace{\frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\{||z_i^0 - T\,\hat{z}_{mid}|| > \frac{\epsilon}{3}\}}}_{(1)} \\
&\qquad + \underbrace{\frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\{||T\,\hat{z}_{mid} - T[\frac{1}{A}\sum_{a=1}^A \hat{c}^a]|| - ||z_{mid}^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| > \frac{\epsilon}{3}\}}}_{(2)} \\
&\qquad + \underbrace{\frac{1}{|\hat{\mathcal{C}}|} \sum_{i \in \hat{\mathcal{C}}} 1_{\{||T[\frac{1}{A}\sum_{a=1}^A \hat{c}^a] - \frac{1}{A}\sum_{a=1}^A c^{a,0}|| > \frac{\epsilon}{3}\}}.}_{(3)}
\end{aligned}
$$

I conclude from Theorem 1 that (1) is an $o_P(1)$ term. For (2) and (3), I exploit that after using the isometry T^{-1} a distance stays the same. Hence, I can immediately conclude from Corollary 6 that (3) is also an $o_P(1)$ term.
It remains to show that $\mathbb{P}\Big[\Big\{\Big|||\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^A \hat{c}^a|| - ||z_{mid}^0 - \frac{1}{A}\sum_{a=1}^A c^{a,0}||\Big| > \frac{\epsilon}{3}\Big\}\Big] \to 0.$

I can derive

$$
\begin{aligned}
&\mathbb{P}\Big[\Big\{\Big|||\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}|| - ||z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}c^{a,0}||\Big| > \frac{\epsilon}{3}\Big\}\Big] \\
&\quad = \mathbb{P}\Big[\Big\{\Big|||\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}|| - ||z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}c^{a,0}||\Big| > \frac{\epsilon}{3}\Big\} \cap \Big\{||T^{-1}(\frac{1}{A}\sum_{a=1}^{A}c^{a,0}) - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}|| \geq \frac{\epsilon}{3}\Big\}\Big] \\
&\qquad + \mathbb{P}\Big[\Big\{\Big|||\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}|| - ||z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}c^{a,0}||\Big| > \frac{\epsilon}{3}\Big\} \cap \Big\{||T^{-1}(\frac{1}{A}\sum_{a=1}^{A}c^{a,0}) - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}|| < \frac{\epsilon}{3}\Big\}\Big].
\end{aligned}
$$

The first term goes to zero by Corollary 6. For the second term, I can make the following arguments.
Assume (otherwise the same arguments hold)

$$\Big\|z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big\| < \Big\|\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\|.$$

Therefore, there exists an $\alpha > \frac{\epsilon}{3}$ such that

$$\Big\|z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big\| = \Big\|\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\| - \alpha.$$

If $T^{-1}z^{0}_{mid} = \hat{z}_{mid}$, the statement of Theorem 7 can be derived as an immediate consequence of Corollary 6. Thus, I assume $T^{-1}z^{0}_{mid} \neq \hat{z}_{mid}$. But then I can deduce

$$
\begin{aligned}
\Big\|T^{-1}\,z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\| &= \Big\|T^{-1}\,z^{0}_{mid} - T^{-1}\Big(\frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big) + T^{-1}\Big(\frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big) - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\| \\
&\leq \Big\|T^{-1}\,z^{0}_{mid} - T^{-1}\Big(\frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big)\Big\| + \Big\|T^{-1}\Big(\frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big) - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\| \\
&\leq \Big\|z^{0}_{mid} - \frac{1}{A}\sum_{a=1}^{A}c^{a,0}\Big\| + \frac{\epsilon}{3} \\
&= \Big\|\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\| - \alpha + \frac{\epsilon}{3} \\
&< \Big\|\hat{z}_{mid} - \frac{1}{A}\sum_{a=1}^{A}\hat{c}^{a}\Big\|.
\end{aligned}
$$

This is in contradiction to $\hat{z}_{mid}$ being the minimizer of the distance to the arithmetic mean of the cluster midpoints. Hence, also term (2) goes to zero in probability.

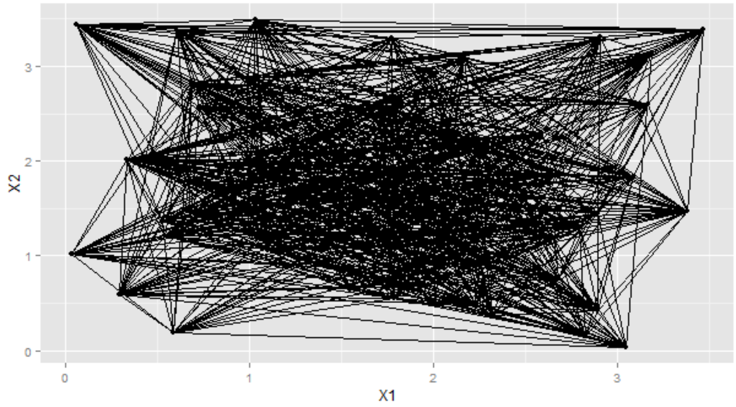


Figure 13: Illustration of the network discussed in Simulation 4.3

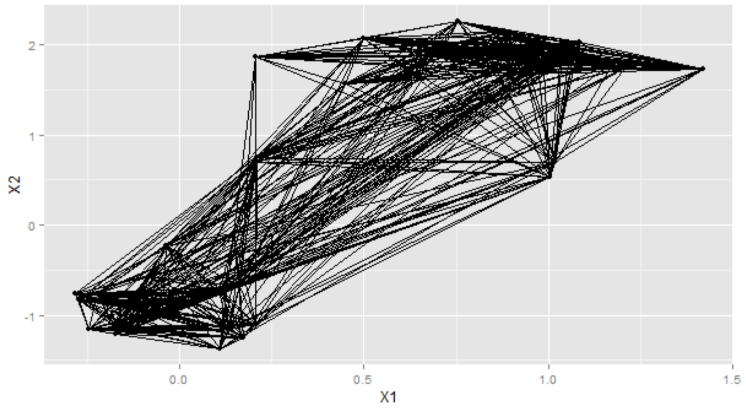


Figure 14: Illustration of estimation of the network discussed in Simulation 4.3

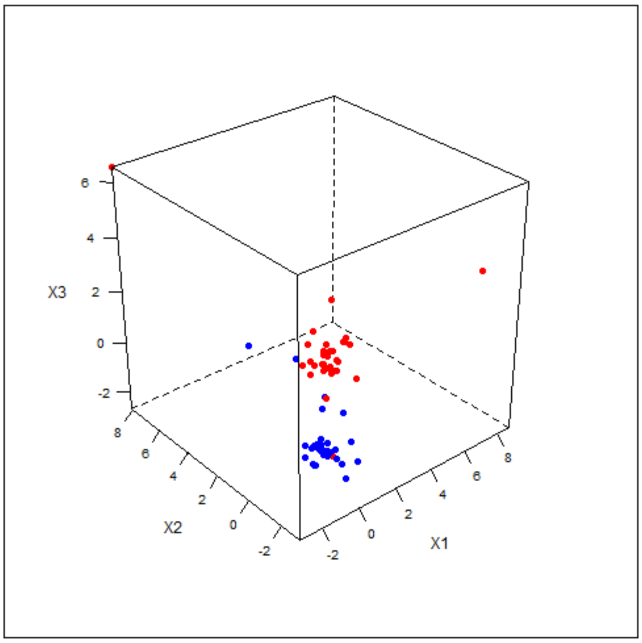


Figure 15: Estimated Positions for Political Blogs (dim)

D. Tables

	$\theta = 0.10$						$\theta = 0.05$					
C_h	0.50	0.75	1.00	1.25	1.50	1.75	0.50	0.75	1.00	1.25	1.50	1.75
null												
$C_g = 0.50$	6.7	5.7	5.8	8.9	7.0	7.9	2.7	2.6	1.9	4.6	3.3	2.8
$C_g = 0.75$	9.2	6.4	8.2	6.5	6.4	7.0	4.6	2.0	3.2	2.8	3.2	2.8
$C_g = 1.00$	6.4	6.7	8.1	6.8	8.8	7.1	2.2	2.9	2.9	3.1	3.2	2.8
alternative 1												
$C_g = 0.50$	65.8	65.8	67.7	63.8	65.3	65.7	50.5	49.9	53.2	47.5	50.6	50.8
$C_g = 0.75$	65.1	65.8	64.8	65.3	65.8	65.9	49.7	47.7	49.9	49.5	50.1	52.3
$C_g = 1.00$	66.3	65.0	66.4	67.9	64.8	66.5	50.4	51.2	50.3	51.1	50.9	49.2
alternative 2												
$C_g = 0.50$	82.4	79.9	80.2	80.5	81.6	78.0	67.9	66.8	68.4	67.3	68.6	65.5
$C_g = 0.75$	79.2	81.0	79.9	80.6	80.4	79.8	66.1	68.3	68.0	68.2	66.5	65.8
$C_g = 1.00$	80.9	81.4	80.1	80.3	80.3	78.2	68.4	67.5	66.1	66.9	64.0	64.7
alternative 3												
$C_g = 0.50$	6.9	8.1	8.8	7.7	5.0	6.7	2.3	3.9	3.3	4.2	1.8	3.2
$C_g = 0.75$	7.2	8.1	6.8	7.4	6.7	6.9	2.9	2.6	3.7	3.9	2.1	3.0
$C_g = 1.00$	7.7	8.0	6.2	6.7	7.8	7.1	2.6	3.3	3.1	2.3	3.5	2.6
alternative 4												
$C_g = 0.50$	15.0	10.5	15.1	14.0	13.1	12.2	7.0	4.8	6.5	5.7	7.0	6.6
$C_g = 0.75$	12.5	13.9	13.8	12.9	13.6	13.3	5.2	6.2	7.0	7.0	6.3	5.9
$C_g = 1.00$	10.0	15.7	14.1	15.7	11.5	14.2	4.2	6.9	7.4	9.5	4.7	6.7
alternative 5												
$C_g = 0.50$	12.0	12.4	15.5	13.5	14.2	13.2	5.7	4.6	7.4	4.9	5.8	6.0
$C_g = 0.75$	13.4	14.5	12.5	12.1	12.0	11.1	6.0	6.9	5.7	4.2	5.3	5.7
$C_g = 1.00$	12.2	14.3	13.1	12.6	12.9	12.2	5.3	5.8	5.7	5.9	6.4	6.2
alternative 6												
$C_g = 0.50$	22.5	23.0	22.9	24.0	21.6	23.1	12.3	12.4	11.2	14.3	10.7	12.8
$C_g = 0.75$	23.3	20.6	25.3	23.5	23.3	20.0	12.5	11.4	13.2	12.0	13.5	12.1
$C_g = 1.00$	22.0	22.0	20.9	25.7	24.0	20.8	11.7	11.6	9.9	13.4	12.7	9.9

Tabelle 3: Simulation. Empirical rejection probabilities in percentage points under nominal level θ. Sample size is $n = 200$.

			D		Y	
Race	Z	n	mean	sd	mean	sd
black	0	787	0.19949	0.3999	0.8183	0.3858
	1	67	0.26866	0.4466	0.6269	0.4873
hispanic	0	549	0.18033	0.3848	0.7687	0.4221
	1	36	0.27778	0.4543	0.5278	0.5063
white	0	1394	0.07389	0.2617	0.8479	0.3592
	1	77	0.20779	0.4084	0.6234	0.4877

Tabelle 4: Teenage child bearing (D) and high-school graduation (Y).

	g	h	T_n	$\underline{x}_1$	$\bar{x}_1$	$\underline{x}_2$	$\bar{x}_2$	$\underline{x}_3$	$\bar{x}_3$	$P(>T_n)$	test result
1	1.00	0.15	0.086	0.03	0.32	0.01	0.21	0.05	0.46	0.225	no rejection
2	1.50	0.15	0.053	0.03	0.27	0.03	0.18	0.05	0.41	0.224	no rejection
3	2.00	0.15	0.084	0.03	0.21	0.02	0.15	0.05	0.36	0.059	*
4	2.50	0.15	0.054	0.04	0.19	0.02	0.13	0.11	0.27	0.012	**
5	3.00	0.15	0.022	0.02	0.18	0.05	0.11	0.13	0.19	0.092	*
6	1.00	0.20	0.064	0.03	0.32	0.01	0.21	0.05	0.46	0.060	*
7	1.50	0.20	0.042	0.03	0.27	0.03	0.18	0.05	0.41	0.084	*
8	2.00	0.20	0.067	0.03	0.21	0.02	0.15	0.05	0.36	0.010	**
9	2.50	0.20	0.043	0.04	0.19	0.02	0.13	0.11	0.27	0.010	**
10	3.00	0.20	0.019	0.02	0.18	0.05	0.11	0.13	0.19	0.083	*
11	1.00	0.25	0.045	0.03	0.32	0.01	0.21	0.05	0.46	0.012	**
12	1.50	0.25	0.037	0.03	0.27	0.03	0.18	0.05	0.41	0.036	**
13	2.00	0.25	0.051	0.03	0.21	0.02	0.15	0.05	0.36	0.008	***
14	2.50	0.25	0.035	0.04	0.19	0.02	0.13	0.11	0.27	0.025	**
15	3.00	0.25	0.017	0.02	0.18	0.05	0.11	0.13	0.19	0.090	*
16	1.00	0.30	0.040	0.03	0.32	0.01	0.21	0.05	0.46	0.010	**
17	1.50	0.30	0.035	0.03	0.27	0.03	0.18	0.05	0.41	0.036	**
18	2.00	0.30	0.044	0.03	0.21	0.02	0.15	0.05	0.36	0.021	**
19	2.50	0.30	0.030	0.04	0.19	0.02	0.13	0.11	0.27	0.022	**
20	3.00	0.30	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.080	*
21	1.00	0.35	0.039	0.03	0.32	0.01	0.21	0.05	0.46	0.005	***
22	1.50	0.35	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.024	**
23	2.00	0.35	0.041	0.03	0.21	0.02	0.15	0.05	0.36	0.014	**
24	2.50	0.35	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.018	**
25	3.00	0.35	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.064	*
26	1.00	0.40	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.003	***
27	1.50	0.40	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.021	**
28	2.00	0.40	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.007	***
29	2.50	0.40	0.028	0.04	0.19	0.02	0.13	0.11	0.27	0.011	**
30	3.00	0.40	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.064	*
31	1.00	0.50	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.003	***
32	1.50	0.50	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.012	**
33	2.00	0.50	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.012	**
34	2.50	0.50	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.005	***
35	3.00	0.50	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.065	*

Tabelle 5: Test results for varying bandwidths and $c_\delta = 0.050$. (*) reject at 0.10 level, (**) reject at 0.05 level, (***) reject at 0.01 level.

	g	h	T_n	$\underline{x}_1$	$\bar{x}_1$	$\underline{x}_2$	$\bar{x}_2$	$\underline{x}_3$	$\bar{x}_3$	$P(>T_n)$	test result
1	1.00	0.15	0.057	0.06	0.29	0.03	0.18	0.07	0.44	0.170	no rejection
2	1.50	0.15	0.033	0.06	0.24	0.05	0.15	0.08	0.39	0.208	no rejection
3	2.00	0.15	0.066	0.06	0.19	0.04	0.13	0.07	0.33	0.042	**
4	2.50	0.15	0.037	0.06	0.16	0.04	0.10	0.13	0.25	0.011	**
5	3.00	0.15	0.009	0.04	0.16	0.07	0.09	0.16	0.16	0.114	no rejection
6	1.00	0.20	0.041	0.06	0.29	0.03	0.18	0.07	0.44	0.038	**
7	1.50	0.20	0.028	0.06	0.24	0.05	0.15	0.08	0.39	0.108	no rejection
8	2.00	0.20	0.048	0.06	0.19	0.04	0.13	0.07	0.33	0.009	***
9	2.50	0.20	0.029	0.06	0.16	0.04	0.10	0.13	0.25	0.014	**
10	3.00	0.20	0.009	0.04	0.16	0.07	0.09	0.16	0.16	0.101	no rejection
11	1.00	0.25	0.033	0.06	0.29	0.03	0.18	0.07	0.44	0.011	**
12	1.50	0.25	0.025	0.06	0.24	0.05	0.15	0.08	0.39	0.044	**
13	2.00	0.25	0.034	0.06	0.19	0.04	0.13	0.07	0.33	0.012	**
14	2.50	0.25	0.023	0.06	0.16	0.04	0.10	0.13	0.25	0.020	**
15	3.00	0.25	0.008	0.04	0.16	0.07	0.09	0.16	0.16	0.113	no rejection
16	1.00	0.30	0.031	0.06	0.29	0.03	0.18	0.07	0.44	0.010	**
17	1.50	0.30	0.024	0.06	0.24	0.05	0.15	0.08	0.39	0.036	**
18	2.00	0.30	0.031	0.06	0.19	0.04	0.13	0.07	0.33	0.015	**
19	2.50	0.30	0.020	0.06	0.16	0.04	0.10	0.13	0.25	0.023	**
20	3.00	0.30	0.007	0.04	0.16	0.07	0.09	0.16	0.16	0.138	no rejection
21	1.00	0.35	0.030	0.06	0.29	0.03	0.18	0.07	0.44	0.005	***
22	1.50	0.35	0.024	0.06	0.24	0.05	0.15	0.08	0.39	0.024	**
23	2.00	0.35	0.029	0.06	0.19	0.04	0.13	0.07	0.33	0.017	**
24	2.50	0.35	0.018	0.06	0.16	0.04	0.10	0.13	0.25	0.013	**
25	3.00	0.35	0.007	0.04	0.16	0.07	0.09	0.16	0.16	0.124	no rejection
26	1.00	0.40	0.030	0.06	0.29	0.03	0.18	0.07	0.44	0.008	***
27	1.50	0.40	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.020	**
28	2.00	0.40	0.029	0.06	0.19	0.04	0.13	0.07	0.33	0.016	**
29	2.50	0.40	0.028	0.04	0.19	0.02	0.13	0.11	0.27	0.005	***
30	3.00	0.40	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.076	*
31	1.00	0.50	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.001	***
32	1.50	0.50	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.012	**
33	2.00	0.50	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.012	**
34	2.50	0.50	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.010	**
35	3.00	0.50	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.062	*

Tabelle 6: Test results for varying bandwidths and $c_\delta = 0.075$. (*) reject at 0.10 level, (**) reject at 0.05 level, (***) reject at 0.01 level.

References

Abadie, Alberto (2003). “Semiparametric instrumental variable estimation of treatment response models”. In: *Journal of Econometrics* 113.2, pp. 231–263.

Adamic, Lada A and Natalie Glance (2005). “The political blogosphere and the 2004 US election: divided they blog”. In: *Proceedings of the 3rd international workshop on Link discovery.* ACM, pp. 36–43.

Advani, Arun and Bansi Malde (2014). “Empirical methods for networks data: Social effects, network formation and measurement error”. In: *Unpublished manuscript, University College London.*

Airoldi, Edoardo M, Thiago B Costa, and Stanley H Chan (2013). “Stochastic blockmodel approximation of a graphon: Theory and consistent estimation”. In: *Advances in Neural Information Processing Systems*, pp. 692–700.

Airoldi, Edoardo M et al. (2009). “Mixed membership stochastic blockmodels”. In: *Advances in Neural Information Processing Systems*, pp. 33–40.

Amini, Arash A et al. (2013). “Pseudo-likelihood methods for community detection in large sparse networks”. In: *The Annals of Statistics* 41.4, pp. 2097–2122.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). “Identification of Causal Effects Using Instrumental Variables”. In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Anthonisse, Jac M (1971). “The rush in a directed graph”. In: *Stichting Mathematisch Centrum. Mathematische Besliskunde* BN 9/71, pp. 1–10.

Balke, Alexander and Judea Pearl (1997). “Bounds on treatment effects from studies with imperfect compliance”. In: *Journal of the American Statistical Association* 92.439, pp. 1171–1176.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou (2006). “Who’s who in networks. wanted: the key player”. In: *Econometrica* 74.5, pp. 1403–1417.

Banerjee, Abhijit et al. (2013). “The diffusion of microfinance”. In: *Science* 341.6144.

Bickel, Peter et al. (2013). “Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels”. In: *The Annals of Statistics* 41.4, pp. 1922–1943.

Bickel, Peter J and Aiyou Chen (2009). “A nonparametric view of network models and Newman–Girvan and other modularities”. In: *Proceedings of the National Academy of Sciences* 106.50, pp. 21068–21073.

Bickel, Peter J, Aiyou Chen, Elizaveta Levina, et al. (2011). “The method of moments and degree distributions for network models”. In: *The Annals of Statistics* 39.5, pp. 2280–2301.

Blume, Lawrence E et al. (2010). “Identification of social interactions”. In: *Available at SSRN 1660002.*

Bonacich, Phillip (1987). “Power and centrality: A family of measures”. In: *American journal of sociology*, pp. 1170–1182.

Bornschier, Simon (2010). “The new cultural divide and the two-dimensional political space in Western Europe”. In: *West European Politics* 33.3, pp. 419–444.

Carneiro, Pedro, James Heckman, and Edward Vytlacil (2011). “Estimating Marginal Returns to Education”. In: *American Economic Review* 101.6, pp. 2754–2781.

Carneiro, Pedro and Sokbae Lee (2009). “Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality”. In: *Journal of Econometrics* 149.2, pp. 191–208.

Chan, Stanley H and Edoardo M Airoldi (2014). "A Consistent Histogram Estimator for Exchangeable Graph Models". In: *arXiv preprint arXiv:1402.1888*.

Chatterjee, Sourav, Persi Diaconis, et al. (2013). "Estimating and understanding exponential random graph models". In: *The Annals of Statistics* 41.5, pp. 2428–2461.

Chernozhukov, Victor, Sokbae Lee, and Adam M Rosen (2013). "Intersection bounds: estimation and inference". In: *Econometrica* 81.2, pp. 667–737.

Choi, David S, Patrick J Wolfe, and Edoardo M Airoldi (2012). "Stochastic blockmodels with a growing number of classes". In: *Biometrika*, asr053.

Christakis, Nicholas A et al. (2010). *An empirical model for strategic network formation.* Tech. rep. National Bureau of Economic Research.

Chung, Fan RK and Linyuan Lu (2006). *Complex graphs and networks.* Vol. 107. American mathematical society Providence.

de Jong, Peter (1987). "A central limit theorem for generalized quadratic forms". In: *Probability Theory and Related Fields* 75.2, pp. 261–277.

Delgado, Miguel A (1993). "Testing the equality of nonparametric regression curves". In: *Statistics & probability letters* 17.3, pp. 199–204.

Dette, Holger and Natalie Neumeyer (2001). "Nonparametric analysis of covariance". In: *the Annals of Statistics* 29.5, pp. 1361–1400.

Doreian, Patrick, Vladimir Batagelj, and Anuska Ferligoj (2005). *Generalized blockmodeling.* 25. Cambridge University Press.

Fernandez-Val, Ivan and Josh Angrist (2013). "ExtrapoLATE-ing: External validity and overidentification in the LATE framework". In: Tenth World Congress. Advances in Economics and Econometrics: Theory and Applications 3. Econometric Society Monographs.

Fienberg, Stephen E and Stanley S Wasserman (1981). "Categorical data analysis of single sociometric relations". In: *Sociological methodology*, pp. 156–192.

Frank, Ove and David Strauss (1986). "Markov graphs". In: *Journal of the american Statistical association* 81.395, pp. 832–842.

Freeman, Linton C (1977). "A set of measures of centrality based on betweenness". In: *Sociometry*, pp. 35–41.

Frölich, Markus (2007). "Nonparametric IV estimation of local average treatment effects with covariates". In: *Journal of Econometrics* 139.1, pp. 35–75.

Goldenberg, Anna et al. (2010). "A survey of statistical network models". In: *Foundations and Trends® in Machine Learning* 2.2, pp. 129–233.

Gørgens, Tue (2002). "Nonparametric comparison of regression curves by local linear fitting". In: *Statistics & probability letters* 60.1, pp. 81–89.

Hall, Peter and Jeffrey D Hart (1990). "Bootstrap test for difference between means in nonparametric regression". In: *Journal of the American Statistical Association* 85.412, pp. 1039–1049.

Hall, Peter and Joel Horowitz (2012). *A simple bootstrap method for constructing nonparametric confidence bands for functions.* Tech. rep. working paper.

Handcock, Mark S, Adrian E Raftery, and Jeremy M Tantrum (2007). "Model-based clustering for social networks". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.2, pp. 301–354.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". In: *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

Härdle, Wolfgang and Enno Mammen (1993). "Comparing nonparametric versus parametric regression fits". In: *The Annals of Statistics* 21.4, pp. 1926–1947.

Heckman, James, Daniel Schmierer, and Sergio Urzua (2010). “Testing the correlated random coefficient model”. In: *Journal of Econometrics* 158.2, pp. 177–203.

Heckman, James, Sergio Urzua, and Edward Vytlacil (2006). “Understanding instrumental variables in models with essential heterogeneity”. In: *The Review of Economics and Statistics* 88.3, pp. 389–432.

Heckman, James and Edward Vytlacil (2005). “Structural Equations, Treatment Effects, and Econometric Policy Evaluation”. In: *Econometrica*, pp. 669–738.

Heckman, James et al. (1996). “Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method”. In: *Proceedings of the National Academy of Sciences* 93.23, pp. 13416–13420.

– (1998). “Characterizing selection bias using experimental data”. In: *Econometrica: Journal of the Econometric Society* 66.5, pp. 1017–1098.

Hoff, Peter D (2005). “Bilinear mixed-effects models for dyadic data”. In: *Journal of the american Statistical association* 100.469, pp. 286–295.

Hoff, Peter D, Adrian E Raftery, and Mark S Handcock (2002). “Latent space approaches to social network analysis”. In: *Journal of the american Statistical association* 97.460, pp. 1090–1098.

Hoffman, Saul D (1998). “Teenage childbearing is not so bad after all... or is it? A review of the new literature”. In: *Family Planning Perspectives* 30.5, pp. 236–243.

Holland, Paul W and Samuel Leinhardt (1977). “A dynamic model for social networksâĂă”. In: *Journal of Mathematical Sociology* 5.1, pp. 5–20.

Hotz, V Joseph, Susan Williams McElroy, and Seth G Sanders (2005). “Teenage Childbearing and Its Life Cycle Consequences Exploiting a Natural Experiment”. In: *Journal of Human Resources* 40.3, pp. 683–715.

Hotz, V Joseph, Charles H Mullin, and Seth G Sanders (1997). “Bounding causal effects using data from a contaminated natural experiment: analysing the effects of teenage childbearing”. In: *The Review of Economic Studies* 64.4, pp. 575–603.

Huber, Martin and Giovanni Mellace (2014). “Testing instrument validity for LATE identification based on inequality moment constraints”. In: *Review of Economics and Statistics.*

Imbens, Guido W. and Joshua D. Angrist (1994). “Identification and estimation of local average treatment effects”. In: *Econometrica: Journal of the Econometric Society*, pp. 467–475.

Jackson, Matthew O (2010). *Social and economic networks.* Princeton University Press.

Kallenberg, Olav (2005). *Probabilistic symmetries and invariance principles.* Vol. 9. Springer.

Katz, Leo (1953). “A new status index derived from sociometric analysis”. In: *Psychometrika* 18.1, pp. 39–43.

King, Eileen, Jeffrey D Hart, and Thomas E Wehrly (1991). “Testing the equality of two regression curves using linear smoothers”. In: *Statistics & Probability Letters* 12.3, pp. 239–247.

Kitagawa, Toru (2013). “A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model”. Working Paper.

Klepinger, Daniel H, Shelly Lundberg, and Robert D Plotnick (1995). “Adolescent fertility and the educational attainment of young women”. In: *Family planning perspectives*, pp. 23–28.

Kong, Efang, Oliver Linton, and Yingcun Xia (2010). “Uniform bahadur representation for local polynomial estimates of M-regression and its application to the additive model”. In: *Econometric Theory* 26.05, pp. 1529–1564.

Kriesi, Hanspeter et al. (2006). "Globalization and the transformation of the national political space: Six European countries compared". In: *European Journal of Political Research* 45.6, pp. 921–956.

Krivitsky, Pavel N et al. (2009). "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models". In: *Social networks* 31.3, pp. 204–213.

Lee, Ying-Ying (2013). "Partial mean processes with generated regressors: Continuous Treatment Effects and Nonseparable models." Working Paper.

Levine, David I and Gary Painter (2003). "The schooling costs of teenage out-of-wedlock childbearing: analysis with a within-school propensity-score-matching estimator". In: *Review of Economics and Statistics* 85.4, pp. 884–900.

Lorrain, Francois and Harrison C White (1971). "Structural equivalence of individuals in social networks". In: *The Journal of mathematical sociology* 1.1, pp. 49–80.

Mammen, Enno (1993). "Bootstrap and wild bootstrap for high dimensional linear models". In: *The Annals of Statistics*, pp. 255–285.

Mammen, Enno, Christoph Rothe, and Melanie Schienle (2012). "Nonparametric regression with nonparametrically generated covariates". In: *The Annals of Statistics* 40.2, pp. 1132–1170.

Masry, Elias (1996). "Multivariate local polynomial regression for time series: uniform strong consistency and rates". In: *Journal of Time Series Analysis* 17.6, pp. 571–599.

Mele, Angelo (2013). "A structural model of segregation in social networks". In: *Available at SSRN 2294957.*

Miller, Amalia R (2011). "The effects of motherhood timing on career path". In: *Journal of Population Economics* 24.3, pp. 1071–1100.

Mourifié, Ismael and Yuanyuan Wan (2014). "Testing LATE Assumptions". In: *Available at SSRN.*

Neumeyer, Natalie and Holger Dette (2003). "Nonparametric comparison of regression curves: an empirical process approach". In: *The Annals of Statistics* 31.3, pp. 880–920.

Newman, Mark EJ (2006). "Finding community structure in networks using the eigenvectors of matrices". In: *Physical review E* 74.3, p. 036104.

Nowicki, Krzysztof and Tom A B Snijders (2001). "Estimation and prediction for stochastic blockstructures". In: *Journal of the American Statistical Association* 96.455, pp. 1077–1087.

Paula, Aureo de, Seth Richards-Shubik, and Elie Tamer (2014). *Identification of Preferences in Network Formation Games.* Tech. rep. Mimeo, Carnegie Mellon University.

Reinhold, Steffen (2007). "Essays in demographic Economics". PhD thesis. John Hopkins University.

Ribar, David C (1994). "Teenage fertility and high school completion". In: *The Review of Economics and Statistics*, pp. 413–424.

Rohe, Karl, Sourav Chatterjee, Bin Yu, et al. (2011). "Spectral clustering and the high-dimensional stochastic blockmodel". In: *The Annals of Statistics* 39.4, pp. 1878–1915.

Ruppert, David and Matthew P Wand (1994). "Multivariate locally weighted least squares regression". In: *The Annals of Statistics*, pp. 1346–1370.

Sargan, John D (1958). "The estimation of economic relationships using instrumental variables". In: *Econometrica: Journal of the Econometric Society*, pp. 393–415.

Shalizi, Cosma Rohilla, Alessandro Rinaldo, et al. (2013). "Consistency under sampling of exponential random graph models". In: *The Annals of Statistics* 41.2, pp. 508–535.

Snijders, Tom AB and Krzysztof Nowicki (1997). "Estimation and prediction for stochastic blockmodels for graphs with latent block structure". In: *Journal of classification* 14.1, pp. 75–100.

Sussman, Daniel L et al. (2012). "A consistent adjacency spectral embedding for stochastic blockmodel graphs". In: *Journal of the American Statistical Association* 107.499, pp. 1119–1128.

Tang, Minh, Daniel L Sussman, Carey E Priebe, et al. (2013). "Universally consistent vertex classification for latent positions graphs". In: *The Annals of Statistics* 41.3, pp. 1406–1430.

Tsallis, Constantino and Daniel A Stariolo (1996). "Generalized simulated annealing". In: *Physica A: Statistical Mechanics and its Applications* 233.1, pp. 395–406.

Vaart, Aad W Van der and Jon A Wellner (1996). *Weak Convergence and Empirical Processes.* Springer.

Vytlacil, Edward (2002). "Independence, monotonicity, and latent index models: An equivalence result". In: *Econometrica* 70.1, pp. 331–341.

Wang, Yuchung J and George Y Wong (1987). "Stochastic blockmodels for directed graphs". In: *Journal of the American Statistical Association* 82.397, pp. 8–19.

Ward, Michael D, Randolph M Siverson, and Xun Cao (2007). "Disputes, democracies, and dependencies: A reexamination of the Kantian peace". In: *American Journal of Political Science* 51.3, pp. 583–601.

Wasserman, Stanley (1994). *Social network analysis: Methods and applications.* Vol. 8. Cambridge university press.

Wasserman, Stanley and Philippa Pattison (1996). "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp". In: *Psychometrika* 61.3, pp. 401–425.

Wolfe, Patrick J and Sofia C Olhede (2013). "Nonparametric graphon estimation". In: *arXiv preprint arXiv:1309.5936*.

Yu, P and H Van de Sompel (1965). "Networks of scientific papers". In: *Science* 169, pp. 510–515.

Zhao, Yunpeng, Elizaveta Levina, and Ji Zhu (2012). "Consistency of community detection in networks under degree-corrected stochastic block models". In: *The Annals of Statistics* 40.4, pp. 2266–2292.

Florian Sarnetzki

Curriculum Vitae

EDUCATION

PhD in Economics (expected May 2015) 2010-2015

University of Mannheim

Dissertation: Econometric Analysis of Heterogeneous Treatment and Network Models

First Chapter:

Overidentification Test in a Nonparametric Treatment Model with Unobserved Heterogeneity (Reinhard Selten Award 2014)

Second Chapter (Job Market Paper):

A Network Centrality Measure and Consistency of ML-Estimation in a Latent Space Model

Diploma (M.Sc equivalent) in Mathematics (Minor: Economics) 2005-2010

University of Heidelberg

Diploma Thesis: Extensions of the MQ-CAViaR model: including volatilities and values at risk into a GARCH-specification

WORK EXPERIENCE

Research Assistant 2012-present

University of Mannheim

- teaching assistant and lecturer for undergraduate and PhD level courses
- two research papers
- presentation of my research at international conferences and invited seminars (Paris School of Economics, ENSAI, University of Gothenburg, University of Groningen)

SKILLS

Languages	German (mother tongue) English (fluent) French (basic)
Software	R, Microsoft Office, Python, Matlab, LaTeX