SUB CRUCE LUMEN

THE UNIVERSITY
of ADELAIDE

Essays in Econometrics with Applications to Social Networks

by

Robert C. Garrard

A Thesis Submitted in Total
Fulfilment of the Requirements
for the Degree of

Doctor of Philosophy

School of Economics
University of Adelaide
September 7, 2017

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed ________________________ Date 01 / 06 / 2017

Abstract

This thesis is comprised of three self-contained essays on econometrics.

The first paper illustrates how stochastic dominance criteria can be used to rank social networks in terms of efficiency, and develops statistical inference procedures for assessing these criteria. The tests proposed can be viewed as extensions of a Pearson goodness-of-fit test and a studentized maximum modulus test often used to partially rank income distributions and inequality measures. We establish uniform convergence of the empirical size of the tests to the nominal level, and show their consistency under the usual conditions that guarantee the validity of the approximation of a multinomial distribution to a Gaussian distribution. Furthermore, we propose a bootstrap method that enhances the finite-sample properties of the tests. The performance of the tests is illustrated via Monte Carlo experiments and an empirical application to risk sharing networks in rural India.

The second paper considers the problem of testing a hypothesis $H_0 : \beta = \beta_0$ where β is a vector representing the degree distribution of a graph and the sample acquired is an induced subgraph. We propose a novel bootstrap procedure to control the size of a test under the null hypothesis by constructing a graph whose degree distribution conforms to the null hypothesis from which we may draw pseudo-samples in the form of induced subgraphs. We investigate the properties of the bootstrap with a simulation study in which a Wald-type statistic based on a truncated singular value estimator, whose null distribution is approximately chi-square, serves as a benchmark. We then discuss whether this test may be inverted to construct confidence intervals.

The third paper presents a selective review of the Lasso estimator as it applies to econometric inference. We survey key papers addressing properties of the Lasso of interest to the econometrician including conditions for consistency, the asymptotic distribution of the estimator, its ability to be bootstrapped, sample splitting for high dimensional inference, and how it may be used to solve the many instruments problem in instrumental variables regression.

Acknowledgements

I'm immensely grateful to Virginie Masson and Firmin Doko Tchatoka for their supervision, guidance, and friendship. I'd like to thank Sarah Cornell-Farrow and Qazi Haque for sharing ideas and keeping economics interesting; my friends and family for making a stressful time bearable; and the Howling Owl, for being my home away from home. I'd also like to thank two anonymous referees for their comments.

Contents

Preface

Traditional neoclassical economic theory tends to view economic behavior through the lens of centralized markets which are perfectly competitive and obtain a market clearing price and quantity in equilibrium. In order to study a richer set of phenomena, modern theory attempts to relax some of these simplifying assumptions. Perfect competition is frequently relaxed through the introduction of a continuum of monopolistically competitive intermediate goods firms or an increasing returns to scale production technology, and the existence of a market clearing price and quantity may be substituted for search and matching models (such as the Diamond-Mortensen-Pissaredes model of labor search). The study of social networks attempts to generalize economic exchange away from centralized markets to interactions between sparsely connected agents. Much theoretical headway has been made toward understanding how the structural features of social networks can affect the underlying behaviors, but empirical techniques for measuring these features are yet to catch up.

Of particular interest in the network theory literature has been in investigating the compatibility between networks which are stable equilibrium outcomes and networks which are efficient in an aggregate social welfare sense. In the event that the aggregate social welfare function in question is increasing and concave in the degree of each agent in the network, such as for risk sharing networks, then efficiency of one network over another may be determined by demonstrating that the degree distribution of one network second-order stochastically dominates the other. The first article in this thesis contributes to econometric inference regarding whether two observed networks may be ranked by such a stochastic dominance criterion. The key assumption underpinning the success of the results in this paper requires that the networks be sampled in a fashion which guarantees independent draws from the degree distribution of a given network. An example of such a valid sampling method is illustrated in the paper's empirical application. We use a data set of borrowing and lending networks from villages in rural India. The researchers who collected this data did so by randomly selecting households in the village and for each household asking its residents to name all of their connections to other members of the village for a particular set of behaviors. That is, nodes were randomly selected and the degree of each node was observed. This constitutes a valid i.i.d sample from the degree distribution of a network.

However, while this is statistically the most tractable way to handle questions regarding the degree distribution, it is not the only or even the most common method for sampling networks. Take, for example, the webgraph of the internet. Sampling in this fashion would first require a list of every webpage from which we may randomly select a subset, which presents an insurmountable task. The observable internet is currently over 14 billion pages and the Deep Web, constituting pages which have not yet been indexed, is estimated to be at least 400 times larger. Instead, measurment of the internet is done by programs called "web crawlers" which start from a specified page, record the set of hyperlinks on that page, choose one of those hyperlinks at random and follow it to the next page. This method, while feasible from a measurement point of view, renders statistical inference extraordinarily difficult.

The second article in this thesis considers a similarly challenging sampling method called induced subgraph sampling. To sample an induced subgraph, one first randomly selects a subset of nodes and includes in the sample *only* links between sampled nodes. At first this might seem quite wasteful, but it makes sense in terms of things we typically like to measure in networks. The computation of measures of centrality, clustering, cliquishness, average path length, etc, all require that links be only to other members of the sampled network. In this article we investigate how one might test the hypothesis that an induced subgraph sample came from a particular degree distribution. Naturally a first best result would be a method for consistently estimating the degree distribution and constructing valid confidence intervals around the point estimate. Given the highly peculiar nature of the sampling method, a consistent estimator for the degree distribution is still an open question in statistics. This article presents a first pass at the problem of constructing confidence intervals by considering the dual problem of performing a simple hypothesis test. While we are able to offer no asymptotic results due to the intractibility of the sample distribution, we do propose a novel bootstrap procedure to control the size of the hypothesis test which we investigate with a monte carlo experiment.

If the degree distribution itself is not consistently estimable one might ask if at very least the *support* of the degree distribution is consistently estimable. Induced subgraph sampling has the interesting property that the form of the distortion to the degree distribution is relatively easy to characterize. The degree distribution of the induced subgraph is a linear mapping of the true degree distribution into a lower dimensional subspace where the elements of the linear map are a function of the number of nodes in the population and the sample size. This means the problem of estimating the degree distribution may be posed as a linear regression with fixed design. However, since the map projects the degree distribution into a lower dimensional space, the linear regression is high dimensional. That is, there are more unknown parameters than observations.

If we hope to solve the problem of estimating the support of a network's degree distribution based on an induced subgraph sample, this involves determining which parameters in a high dimensional linear regression are equal to zero. This is the realm of the Lasso estimator. Aside from a potential application to network econometrics, the Lasso has reason to be of interest to the economic researcher in general. The Lasso simultaneously estimates a parameter and performs model selection. By setting some of the estimated coefficients exactly to zero, it in effect decides which of a large set of regressors are important for explaining the dependent variable. In recent years both psychology and economics have had large projects attempt to replicate several published results and found that many of the results fail to be reproducible. There are two leading explanations for this replication failure. One is that negative results, in which a paper does not find a statistically significant effect, is unlikely to be accepted by a journal for publication, and so the rate of false discovery is much higher in published works. The other is that through either deliberate or subconscious decisions a researcher will manage to engineer statistical significance through experimenting with which variables do or do not enter the final model; so called 'p-hacking'. If there is any kind of model search in which many models are tried in order to find one which fits the data well, then inferences based on the final model lose any frequency guarantees on which confidence is based. That is, 95% confidence intervals will have less than 95% coverage. The Lasso is a method in which model selection is automated and this model selection is internalized when conducting inferences. The third article presents a selective review of key papers in the development of the Lasso with direct relevance to causal econometric inference.

Statement of Authorship

Title of Paper	**Testing for Stochastic Dominance in Social Networks**	
Publication Status	□ Published □ Submitted for Publication	□ Accepted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details		

Principal Author

Name of Principal Author (Candidate)	Robert Garrard		
Contribution to the Paper			
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	01/06/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Firmin Doko Tchatoka		
Contribution to the Paper	Proof of bootstrap validity. Body of section 3. Parts of introduction.		
Signature		Date	01/06/2017

Name of Co-Author	Virginie Masson		
Contribution to the Paper	Network notation. Tikz images. Body of section 2. Parts of introduction.		
Signature		Date	01/06/2017

Please cut and paste additional co-author panels here as required.

Testing for Stochastic Dominance in Social Networks

Firmin Doko Tchatoka, Robert Garrard, Virginie Masson

School of Economics, The University of Adelaide

Abstract

This paper illustrates how stochastic dominance criteria can be used to rank social networks in terms of efficiency, and develops statistical inference procedures for assessing these criteria. The tests proposed can be viewed as extensions of a Pearson goodness-of-fit test and a studentized maximum modulus test often used to partially rank income distributions and inequality measures. We establish uniform convergence of the empirical size of the tests to the nominal level, and show their consistency under the usual conditions that guarantee the validity of the approximation of a multinomial distribution to a Gaussian distribution. Furthermore, we propose a bootstrap method that enhances the finite-sample properties of the tests. The performance of the tests is illustrated via Monte Carlo experiments and an empirical application to risk sharing networks in rural India.

Keywords: Networks, Tests of stochastic dominance, Bootstrap, Uniform convergence

JEL: C12, C13, C36

1. Introduction

This paper considers the problem of assessing stochastic dominance criteria in network theory. Many economic and social interactions involve network relationships, and the role that networks play in determining economic outcomes– such as trade and exchange of goods in non-centralized markets (e.g., Tesfatsion (1997)), provision of mutual insurance in developing countries (e.g., Fafchamps and Lund (2003)), and job search (e.g., Calvo-Armengol (2004))– is now recognized. Recent statistical and econometric studies in network theory have often focused on the estimation of network relationships,[1] and the identification of peer effects.[2] Statistical methods for understanding how individual incentives to form networks align with social efficiency are yet to be developed.

This paper illustrates how *stochastic dominance* criteria can be used to rank networks in terms of social efficiency, and proposes a nonparametric procedure for assessing these criteria. Often, standard measures– such as the Gini-coefficient or Lorenz curves– are used to rank income and poverty distributions in terms of social efficiency. However, in addition to being relative measures,[3] two income or poverty distributions such that one second-order statistically dominates the other may result in a same value of these measures. For theses reasons, stochastic dominance criteria are usually preferred to provide a partial ordering of inequality and poverty measures (e.g., Atkinson (1987) and Anderson (1996)), and the concept, as well as its connection to social welfare theory, now extends to network theory (e.g., Goyal (2012) and Jackson et al. (2008)). To illustrate how the

[1] See Chandrasekhar (2015), Leung (2015), Banerjee et al. (2013), Liu (2013), Bickel et al. (2011), and Bickel and Chen (2009) among others.

[2] See Hsieh and Lee (2016), Blume et al. (2015), Bursztyn et al. (2014), Goldsmith-Pinkham and Imbens (2013), Jackson (2014), Graham (2014), and Aliprantis and Richter (2013) among others.

[3] For example, changing income inequality, measured by Gini-coefficients, can be due to structural changes in a society such as aging populations, emigration,immigration, etc.

stochastic dominance criteria could provide a partial ordering of networks, let $N = \{1, 2, \ldots, n\}$ be a finite set of n agents and $G(N)$ be the set of networks on N. Let $\mathbb{W}(d_g)$ denote the aggregate social welfare function of network $g \in G(N)$, where $d_g = (d_{g\cdot 1}, \ldots, d_{g\cdot n})'$ and $d_{g\cdot i}$ is the degree of agent $i \in N$ in g. Following Goyal (2012, Section 7.4), network $g \in G(N)$ is said to be *socially efficient* if $\mathbb{W}(d_g) \geqslant \mathbb{W}(d_{g'})$ for all $g' \in G(N)$. Therefore, if $\mathbb{W}(d_g)$ is a nondecreasing and strictly concave function of $d_{g\cdot i}$ for all $i \in N$, then second-order stochastic dominance between the degree distributions of two networks g and g' in $G(N)$ is equivalent to dominance between $\mathbb{W}(d_g)$ and $\mathbb{W}(d_{g'})$ in the same direction (e.g., Rothschild and Stiglitz (1970)). [4] Therefore, the stochastic dominance criteria provide a partial ordering of the elements of $G(N)$ in terms of social efficiency in this setting, and developing statistical methods to establish this ordering from the observed network relationships can be of great interest in social science.

Tests similar to that of Pearson (1900) are often used for assessing stochastic dominance hypotheses in the literature on inequality and poverty measures,[5] but to the best of our knowledge, this study is the first to focus on extending these procedures to network theory. Anderson (1996) suggests a combination of Pearson-type and studentized maximum modulus (SMM) tests[6] in a single decision rule for assessing stochastic dominance of income

[4] A sufficient condition for a social welfare function with such properties is for the welfare function to be additive in individual utilities, $\mathbb{W} = \sum_{i=1}^{n} u_i$, with individual utility functions of the form $u_i = f(d_i) - \sum_{j \in \mathcal{N}(i)} g(d_j)$, with $f(d_i)$ being an increasing and concave benefit in own degree and $g(d_j)$ being the disutility in the degree of neighbor j such that the function $xg(x)$ is quasi-convex; for example Calv-Armengol (2004), Bramoulle and Kranton (2007), Choi et al. (2013).

[5] For example, see McFadden (1989), Anderson (1996), Davidson and Duclos (2000), Barrett and Donald (2003), Linton et al. (2005), and Barrett et al. (2014).

[6] See Stoline and Ury (1979) for the tabulation of the critical values of the SMM statistics.

distributions. His methodology is nonetheless not directly applicable in the context of networks for the following reasons. *First*, both tests are derived in his framework under the assumption that the samples are independent. Although this may be reasonable in the literature on income distributions and poverty measures, it is less likely to be the case in network theory, as it excludes interesting situations where networks' populations overlap. For example, when comparing risk sharing networks formed by men and women within a village (or community), it is reasonable to assume that the two networks are independent across households, while the correlation between the two networks is likely high within households. [7] *Second*, partitioning of samples into classes is usually required to implement a Pearson-type test, and it is well documented that such a partitioning has an influence on the properties (size and power) of the resulting test.[8] In the case where the samples are drawn from a continuous distribution, Mann and Wald (1942) and Williams (1950) propose rules of thumb to select the number of classes and the lengths of subsequent intervals such that the resulting test is unbiased. These optimal rules are usually obtained by equalizing cell probabilities under the null whilst maintaining an expected cell frequency of at least 5 (e.g., Anderson (1996)). The main difficulty in extending Mann and Wald (1942) and Williams (1950) rules of thumb to the context of networks resides in the finite and discrete nature of the range of a network's degree distribution.

Our contribution in this paper is threefold. *First*, we propose an adjustment to Mann and Wald (1942) and Williams (1950) rules of thumb that

[7] Strictly speaking, we are considering correlation between the random variables which generate the degree of each node in the network data generating process.

[8] See Hotelling (1930), Mann and Wald (1942), Gumbel (1943), Williams (1950), Cochran (1952), and Schorr (1974) among others.

applies to the context of networks. We show how the optimal choice of the number of classes can be approximated through a careful analysis of the empirical histogram of the degree distributions of the networks. *Second*, we propose a generalization of the Pearson- and SMM-type statistics in Anderson (1996) that are valid even when the samples are correlated, thus applicable to the context of network theory. Our statistics differ from that of Anderson (1996) and prior literature not only through the correction to account for the correlation between the degree distributions of the networks, but also their direct dependence on partitioning into classes. We show that a combination of the two modified statistics into a single decision rule is necessary to inform us on whether *stochastic dominance* holds or not, once equality between the degree distributions of the networks is rejected. As the modified statistics depend on partitioning into classes, controlling the size of the resulting tests uniformly over the set of all *admissible partitions*[9] is important for the asymptotic results to give a good approximation of the empirical size to the nominal level. *Finally*, we provide a bootstrap procedure that improves the finite-sample performance of both the modified Pearson- and SMM-statistics.

We provide an analysis of both the size and power properties of the tests under weaker assumptions than is usually the case in most applications of Pearson's (1900) goodness-of-fit test. On level control, we establish uniform convergence of their empirical size to the nominal level over the set of all *admissible partitions* when the usual asymptotic chi-square and SMM critical values are applied. On power, we show that test consistency

[9] An *admissible partition* is a partition in which the minimum expected number in each cell is at least 5.

holds no matter which *admissible* partition is used. Moreover, we establish uniform consistency of the bootstrap for the two modified Pearson- and SMM-tests irrespective of whether the null hypothesis holds or not. We present a Monte Carlo experiment that confirms our theoretical findings. In particular, while the standard tests sometimes tend to over-reject the null hypothesis if the sample size is small, the bootstrap tests have an overall good performance in such contexts. Finally, using the data set of Jackson et al. (2012) and Banerjee et al. (2012, 2013), we illustrate our theory through an investigation of the households' *risk sharing networks* across 75 villages in rural India. In particular, we focus on both the goods lending and money lending networks, and test gender differences within these networks by applying the tests of stochastic dominance developed. For goods lending, both the standard and bootstrap tests show that the *female network* first- and second-order stochastically dominates the *male network* at the 1% and 5% nominal levels. However, for money lending, we could only find evidence of the first- and second-order dominance of the *female network* at the 5% nominal level. At the 1% nominal level, neither network dominates the other with both the standard and bootstrap tests. These results suggest that women within these villages overall tend to form denser *risk sharing networks* than do men, especially for goods lending.

Throughout this paper, for any vector $x = (x_1, \ldots, x_k)' \in \mathbb{R}^k$, the notation "$x \leqslant 0$" means $x_l \leqslant 0$ for all $l = 1, \ldots, k$, while "$x \nleqslant 0$"(or "$x \ngeqslant 0$") means that there exists l and l' in $\{1, \ldots, k\}$ such that $x_l \geqslant 0$ and $x_{l'} < 0$ or $x_l > 0$ and $x_{l'} \leqslant 0$. Convergence almost surely is symbolized by "$a.s.$", "$\xrightarrow{p}$" stands for convergence in probability, while "$\xrightarrow{d}$" means convergence in distribution. The usual stochastic orders of magnitude are denoted by

$O_p(.)$, $o_p(.)$. $\mathbb{P}[\cdot]$ denotes the relevant probability measure and $\mathbb{E}[\cdot]$ is the expectation operator under $\mathbb{P}[\cdot]$. $\mathbb{P}^*[\cdot]$ is the bootstrap analogue of $\mathbb{P}[\cdot]$, and similarly for $\mathbb{E}^*[\cdot]$. I_q stands for the identity matrix of order q, and for any $q \times q$ matrix A, A^- is the generalized inverse of A. The notation $\mathbf{diag}(A)$ is a $q \times q$ diagonal matrix with diagonal elements the $(l,l)^{th}$ elements of A. $\|U\|$ denotes the usual Euclidian or Frobenius norm for a matrix U. For any set $\mathscr{C}$, $\partial\mathscr{C}$ is the boundary of $\mathscr{C}$ and $(\partial\mathscr{C})^\epsilon$ its ϵ-neighborhood. Finally, $\sup_{\omega\in\Omega}|f(\omega)|$ is the supremum norm on the space of bounded continuous real functions, with topological space Ω.

The remainder of the paper is organised as follows. Section 2 defines the relevant concepts and introduces the dominance criterion. Section 3 formulates the hypotheses tested and presents the basic notations and assumptions used. Section 4 presents the derivation of the statistics and the asymptotic theory developed. Section 5 illustrates the performance of the tests via Monte Carlo experiments. Section 6 provides an empirical illustration of our theoretical results, and Section 7 concludes. Proofs are presented in the appendix.

2. Preliminaries

Before introducing the concept of stochastic dominance in networks (Section 2.2) and formalizing the testing problem of interest (Section 3), we define the basic terminologies and notations used throughout the study.

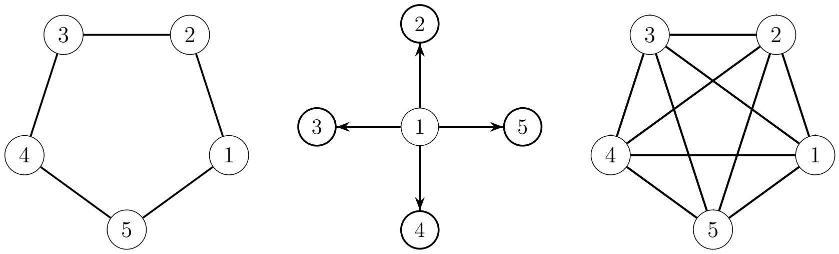
2.1. Networks

Let $N = \{1, 2, \ldots, n\}$ denote a finite set of agents, and $G(N)$ be the set of networks on N. We define a *network* g over N as a pair of nodes and edges describing relationships (or links) between agents $1, 2, \ldots, n$. A network can be represented by a graph whose $n \times n$ adjacency matrix has generic element $g_{ii'}$ satisfying $g_{ii'} = 1$ if there is a directed link from agent i to i', and $g_{ii'} = 0$ otherwise. By convention, we set $g_{ii} = 0$ for all i. The *neighborhood* of agent i is the set of agents with whom i has a directed link in network g, i.e., the set $\mathscr{N}_i(g) = \{i' \in N | g_{ii'} = 1\}$. We refer to the number of agent i's neighbors, $d_{g \cdot i} = card[\mathscr{N}_i(g)]$, as the *degree* of agent i.[10]

The *degree distribution* of network g is a vector $P_g = [\hat{P}_{g \cdot 0}, \ldots, \hat{P}_{g \cdot k}, \ldots, \hat{P}_{g \cdot (n-1)}]'$, where $\hat{P}_{g \cdot k} = card[\{i : d_{g \cdot i} = k\}]/n$ is the proportion of nodes with degree k; thus $\hat{P}_{g \cdot k} \geqslant 0$ for each $k \in \mathscr{R}_n$, $\sum_{k \in \mathscr{R}_n} \hat{P}_{g \cdot k} = 1$, and $\mathscr{R}_n = \{0, 1, \ldots, n-1\}$ is the range of P_g. The empirical *cumulative distribution function* (cdf) of network g is the function $F_g : \mathscr{R}_n \to [0, 1]$ such that $F_g(k) = \sum_0^k \hat{P}_{g \cdot l}$ for all $k \in \mathscr{R}_n$.

Example 1. Figure 1 illustrates three networks with $n = 5$ agents: a "circle" network (Network g), a "directed star" network (Network g'), and a "complete" network (Network g'').

Figure 1: Example of networks



(a) Network g (b) Network g' (c) Network g''

[10] Our definition of a neighborhood considers the *out-degree* of agent i, i.e. the number of links which originate from agent i. However, it can also be defined using the *in-degree* of agent i, in which case, $\mathscr{N}_i(g) = \{i' \in N | g_{i'i} = 1\}$. The choice of the definition depends mainly upon the application considered. For undirected networks, $g_{ii'} = g_{i'i}$ and both definitions coincide.

The characteristics of each network $j \in \{g, g', g''\}$, as per the above terminologies and definitions– *neighborhood*: $\mathscr{N}(j)$, *degree* of agent: $d_{j \cdot i}$, *degree distribution*: P_j, and *empirical cdf*: F_j –are summarized in Table 1.

Table 1: Characteristics of network $j \in \{g, g', g''\}$

characteristics ↓ Network j →	g	g'	g''
$\mathscr{N}_1(j)$	$\{2,5\}$	$\{2,3,4,5\}$	$\{2,3,4,5\}$
$\mathscr{N}_2(j)$	$\{1,3\}$	$\emptyset$	$\{1,3,4,5\}$
$\mathscr{N}_3(j)$	$\{2,4\}$	$\emptyset$	$\{1,2,4,5\}$
$\mathscr{N}_4(j)$	$\{3,5\}$	$\emptyset$	$\{1,2,3,5\}$
$\mathscr{N}_5(j)$	$\{1,4\}$	$\emptyset$	$\{1,2,3,4\}$
$d_{j \cdot 1}$	2	4	4
$d_{j \cdot 2}$	2	0	4
$d_{j \cdot 3}$	2	0	4
$d_{j \cdot 4}$	2	0	4
$d_{j \cdot 5}$	2	0	4
P_j	$(0,0,1,0,0)'$	$(4/5,0,0,0,1/5)'$	$(0,0,0,0,1)'$
F_j	$(0,0,1,1,1)'$	$(4/5,4/5,4/5,4/5,1)'$	$(0,0,0,0,1)'$

2.2. Stochastic Dominance in Networks

Consider the setup described in Section 2.1, and let g and g' denote two networks in $G(N)$ with empirical cdfs F_g and $F_{g'}$, respectively. The first- and second-order[11] stochastic dominance between g and g' are characterized as follows.

Definition 1. (**i**) Network g **first-order stochastically dominates** network g', which we write $g \succ_1 g'$, if $F_g(k) \leqslant F_{g'}(k)\ \forall\ k \in \mathscr{R}_n$, with strict inequality for some k.

[11] The characterization of stochastic dominance can easily be extended to higher-order, but for simplicity we mainly focus on the first- and second-order dominance for the remainder of the paper.

(**ii**) Network g **second-order stochastically dominates** g', which we write $g \succ_2 g'$, if $\sum_{i=0}^{k} F_g(i) \leqslant \sum_{i=0}^{k} F_{g'}(i)\ \forall\ k \in \mathscr{R}_n$, with strict inequality for some k.

We may think of first-order stochastic dominance as describing one network being much more densely connected than another. Given a degree d, if network g has at least as many nodes of degree d than network g', then the average degree in network g will be higher than in g'. Second-order dominance may be thought of in terms of mean-preserving spreads. If two networks have the same average degree, but g has much lower spread around that average than g', then g second-order dominates g'. That is, first-order dominance favors networks where each node has high degree and second-order dominance favors more evenly distributed degrees.

It is straightforward to see from the above characterizations that first-order stochastic dominance implies second-order stochastic dominance, but not the other way around. We now illustrate the two concepts from the example of Section 2.1.

Example 1 (continued)**.** Again, consider the three networks g, g', and g'' of Example 1. From Table 2 below, the pairwise comparisons between the cumulative distributions of these networks show that g'' first-order stochastically dominates both g and g'. Therefore, g'' also second-order stochastically dominates both g and g'. However, as $F_g(1) < F_{g'}(1)$ and $F_g(2) > F_{g'}(2)$, there exists no first-order stochastic dominance between g and g'. Nevertheless, g second-order stochastically dominates g'. This reflects the fact that network g has an average degree at least as high as network g' but a lower dispersion in agents' degrees.

Table 2: Stochastic dominance between networks g, g' and g'' of Example 1

k	0	1	2	3	4
$\hat{p}_{g \cdot k}$	0	0	1	0	0
$F_g(k)$	0	0	1	1	1
$\sum_{i=0}^{k} F_g(i)$	0	0	1	2	3
$\hat{p}_{g' \cdot k}$	0.8	0	0	0	0.2
$F_{g'}(k)$	0.8	0.8	0.8	0.8	1
$\sum_{i=0}^{k} F_{g'}(i)$	0.8	1.6	2.4	3.2	4.2
$\hat{p}_{g'' \cdot k}$	0	0	0	0	1
$F_{g''}(k)$	0	0	0	0	1
$\sum_{i=0}^{k} F_{g''}(i)$	0	0	0	0	1

We now wish to formulate hypotheses for assessing stochastic dominance in social networks from observed real world data.

3. Stochastic Dominance Hypothesis and Assumptions

We first formulate the problem of testing stochastic dominance hypotheses in Section 3.1. Section 3.2 presents the basic notations and assumptions that are used in the paper.

3.1. Hypothesis Formulation

Let g and g' be two networks observed on the same population of n agents, and let F_j denote the empirical cdf associated with the degree distribution P_j of network $j \in \{g,\ g'\}$. Finally, let $\mathbb{N} = \{1, 2, \ldots\}$ be the set of natural integers. Given $m \in \mathbb{N}$, we are interested in assessing which network mth-order *stochastically dominates* the other. [12]From Definition 1, this problem

[12] While definition 1 may be expanded to m-th order stochastic orderings, and the hypothesis below may be tested for any $m \in \mathbb{N}$, it is impractical in most cases to go

can be formulated as a problem of testing the mth-order stochastic dominance between the cdfs F_g and $F_{g'}$, i.e.,

$$H_{0m} : F_g \overset{d}{=} F_{g'} \text{ versus } H_{1m} : F_g \succ_m F_{g'} \;\wedge\; H_{2m} : F_g \overset{d}{\neq} F_{g'}, \qquad (1)$$

where "$\succ_m$" denotes the mth-order stochastic dominance operator, "$\overset{d}{=}$" and "$\overset{d}{\neq}$" symbolize equality and difference in distribution respectively. As can be seen clearly from (1), H_{0m} tests equality between F_g and $F_{g'}$ against: (i) mth-order stochastic dominance (H_{1m}), and (ii) no mth-order dominance (H_{2m}). For example when $m = 2$, H_{02} tests the equality between F_g and $F_{g'}$ against both second-order stochastic dominance (H_{12}) and no second-order dominance (H_{22}). Several statistical procedures exist to assess stochastic dominance hypotheses between two distributions, but to the best of our knowledge, this study is the first to focus on extending these procedures to network theory.

In order to derive a testable formulation of problem (1) from the observed data, as well as test statistics for assessing it, it is useful to first introduce the following notations and assumptions.

3.2. Basic Notations and Assumptions

Let $\{(d_{g\cdot i},\, d_{g'\cdot i})\}_{i=1}^{n}$ be a sample of n observations drawn from the joint distribution of the *degree of agents* in networks g and g'. Let F_g and $F_{g'}$ denote the empirical cdfs of networks g and g' respectively, constructed as in Section 2.1. To build Pearson-type statistics for assessing H_{0m} in (1), we must first partition the range (support) of the *degree distributions* of

beyond $m = 2$ since for such a stochastic ordering to imply a welfare ranking we would have to place more stringent assumptions on the higher order derivatives of the welfare function.

networks g and g' into classes (or class intervals). To do this, we adapt the methodology in Anderson (1996) to the context of social networks.

Let $(d_i)_{i=1}^{2n}$ be the pooled sample of $2n$ observations obtained by stacking the two sub-samples $(d_{g\cdot i})_{i=1}^{n}$ and $(d_{g'\cdot i})_{i=1}^{n}$, and let *Supp(d)* $\subseteq \mathscr{R}_n$ denote the support of the distribution of $(d_i)_{i=1}^{2n}$, where $\mathscr{R}_n = \{0, 1, 2, \ldots, n-1\}$ is the common range of the *degree distributions* of networks g and g'. Note that *Supp(d)* need not be strictly equal to $\mathscr{R}_n$. This is the case for example if $\max_{i,j\in\{g,g'\}} \{d_{j\cdot i}\}_{i=1}^{n} < n-1$. For some fixed $k \in \mathbb{N}$, let $\mathbf{P}_n^{(k)}(\mathbf{I}_1, \ldots, \ \mathbf{I}_k) \equiv \mathbf{P}_n^{(k)}(\mathbf{I}) := \{\mathbf{I}_l\}_{l=1}^{k}$ denote a finite partition of *Supp(d)* into k disjoint sets, i.e.

$$Supp(d) = \bigcup_{1\leqslant l\leqslant k} \mathbf{I}_l : \ \mathbf{I}_l \neq \emptyset, \ \mathbf{I}_l \cap \mathbf{I}_{\tilde{l}} = \emptyset \ \forall \ l \neq \tilde{l}, \tag{2}$$

and define a collection of such partitions by

$$\mathscr{P} = \left\{ \mathbf{P}_n^{(k)}(\mathbf{I}) : \ \mathbf{I} = \{\mathbf{I}_l\}_{l=1}^{k} \text{ satisfies (2)} \right\}. \tag{3}$$

As *Supp(d)* is a discrete finite set, the collection $\mathscr{P}$ contains a finite number of elements (or partitions) for a given k. Until now, we have implicitly assumed that the number k of subsets and the division points between subsets (subsets' cardinality) in (2) are available to the investigator. In practice, one has to choose k as well as the division points between the k resulting subsets, and it is well documented that these choices have an influence on the properties (size and power) of Pearson-type tests. For samples generated from continuous distributions, we have *Supp(d)* $\subseteq \mathbb{R}$ and $\mathbf{I}_l$, $l = 1, 2, \ldots, k$ are compact intervals in (2). In this case, there is a number of seminal papers which provide rules to select k and the lengths of subsequent intervals such that the resulting Pearson-type test is unbiased.

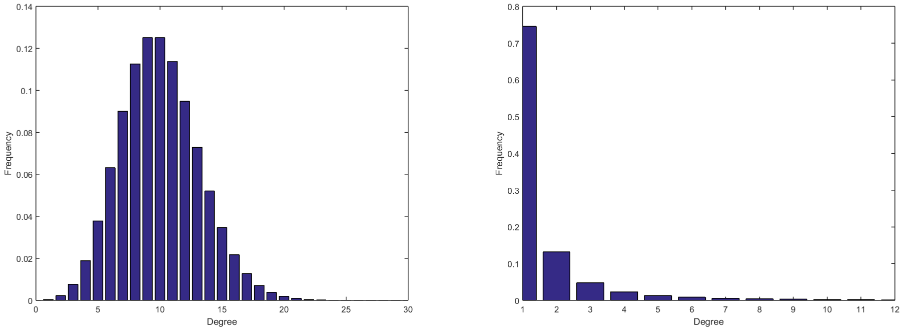
For example, Anderson (1996) suggests that power can be gained by locating partition points at fractiles where it is thought that the two distributions may intersect. Since this information is unknown, the standard advice by Mann and Wald (1942),[13] Gumbel (1943), and Williams (1950), that power is gained by equalizing cell probabilities under the null whilst maintaining an expected cell frequency of at least 5 is usually used in applied work.

The main difficulty in extending Mann and Wald (1942) and Williams (1950) rules of thumb to the context of networks resides in the finite and discrete nature of the range of a network's degree distribution. For example, Figure 2 shows the degree distributions of two commonly used networks: the *Poisson random graph* and the *Scale-free network*. While in theory the range of both distributions is the entire positive integer set $\mathbb{N}$, we see that both distributions are concentrated between: 1–20 (for the *Poisson random graph*), and 1–9 (the *Scale-free network*). Suppose we have a joint sample of $n = 500$ realizations of networks g and g' drawn from a population that follows one of these distributions. For a test at the $\alpha = 5\%$ nominal level ($c = 1.64$), Mann and Wald's (1942) and Williams's (1950) optimal rules of thumb give $k_{_{MW}} = 45$ and $k_{_W} = 23$ respectively. These choices increase to $k_{_{MW}} = 59$ and $k_{_W} = 30$ for a population of $n = 1,000$ agents. However, Figure 2 shows clearly that even a choice of $k \equiv k_{_W} = 23$ in

[13] Mann and Wald (1942) show that the optimal choice of the number of classes is $k := \mathscr{I}nt\left[4\sqrt[5]{\frac{2(n-1)^2}{c^2}}\right]$, where n is the sample size, $\mathscr{I}nt[x]$ is the integer part of any real x, and c is determined so that $\frac{2}{\sqrt{2\pi}}\int_c^\infty e^{-x^2/2}dx$ is equal to the size of the critical region under H_{02}. One criticism of Mann and Wald's (1942) method is that it generates an unnecessarily large number of classes; see Schorr (1974). Williams (1950) shows that halving this number does not substantially decrease the power of Pearson-type tests. Although these rules of thumb are reasonable to follow, it is worth noting that they do not imply that the resulting Pearson-type test is necessarily uniformly powerful against all alternatives; for example, see Cochran (1952).

(2) does not make it possible to equalize cell probabilities under the null whilst maintaining an expected cell frequency of at least 5. Even though this criterion may give a good approximation for *Poisson random graphs* in some instances (for example when λ is large enough), this is likely not the case for *Scale-free networks*. Therefore, adjustments are needed to adapt Mann and Wald's (1942) and Williams's (1950) rules of thumb to the network context. For this purpose, define $k_{max} = \max Supp(d)$. Then, a practical and simple rule of thumb could be to choose $k \leqslant \min[k_{_W}, k_{max}]$ such that Williams (1950) rule of thumb is close to being fulfilled. This can be achieved through a careful analysis of the empirical histogram of the degree distributions such as in Figure 2. For example, if the realizations of networks g and g' are drawn from a *Poisson* population (Figure 2-(a)), both choices: (i) $k = 4$ and $\mathbf{I}_1 = \{1, \ldots, 7\}$, $\mathbf{I}_2 = \{8, 9\}$, $\mathbf{I}_3 = \{10, 11\}$, $\mathbf{I}_4 = \{12, \ldots, 20\}$, and (ii) $k = 4$ and $\mathbf{I}_1 = \{1, \ldots, 9\}$, $\mathbf{I}_2 = \{10\}$, $\mathbf{I}_3 = \{11\}$, $\mathbf{I}_4 = \{12, \ldots, 20\}$, are acceptable. However, the former is closer to the recommendation to equalize cell probabilities than the latter.

Figure 2: The distribution of degrees for Poisson and Scale-free networks



(a) Poisson with parameter $\lambda = 10$

(b) Scale-free with parameter $\gamma = 2.5$

To formally address the threshold of an expected cell frequency of at least 5, we first introduce the following notations and definitions. Let $p_{j \cdot il}, j \in \{g, g'\}$ be the probability that $d_{j \cdot i}$ falls in $\mathbf{I}_l$, and $\hat{p}_{j \cdot l}$ denote the proportion of observations in $(d_{j \cdot i})_{i=1}^n$ which fall in $\mathbf{I}_l$, i.e.

$$p_{j \cdot il} = \mathbb{P}(d_{j \cdot i} \in \mathbf{I}_l), \;\; \hat{p}_{j \cdot l} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(d_{j \cdot i} \in \mathbf{I}_l). \tag{4}$$

If $\{(d_{g \cdot i},\, d_{g' \cdot i})\}_{i=1}^n$ is i.i.d. across i, for given $j \in \{g, g'\}$ and $l \in \{1, \ldots, k\}$, the probabilities $p_{j \cdot il}$ are the same for all i, i.e., $p_{j \cdot il} \equiv p_{j \cdot l}$ for all i and $\hat{p}_{j \cdot l}$ is a consistent estimator of $p_{j \cdot l}$. Then, the *expected numbers* in cell l for network j is given by

$$n_{j \cdot l} := n\hat{p}_{j \cdot l} = \sum_{i=1}^n \mathbb{1}(d_{j \cdot i} \in \mathbf{I}_l). \tag{5}$$

To insure a valid approximation of the *multinomial distribution* to a *multivariate normal distribution*, (2) must also guarantee that the minimum of the $n_{j \cdot l}$'s for all $j \in \{g, g'\}$ and $l = 1, \ldots, k$ is at least 5. This threshold is usually imposed and the absence of a theory to justify its validity has raised some concerns in several seminal papers; e.g., Cochran (1952), Lewis and Burke (1949), and Edwards (1950). Yates (1934) provides a correction for continuity that adjusts the formula for a Pearson-type statistic when this threshold is violated. In this paper, we do not address the issues related to the choice of the minimum expected number in cells. Rather, we consider the collection of all partitions $\mathbf{P}_n^{(k)}(\mathbf{I})$ for which this requirement is satisfied, and we wish to provide tests of stochastic dominance that control the size uniformly over this collection of partitions.

To be more specific, consider the partitions $\mathbf{P}_n^{(k)}(\mathbf{I})$ in (2) such that $n\hat{p}_{j \cdot l} > 5$ for all $j \in \{g, g'\}$ and $l \in \{1, \ldots, k\}$. Let $\mathscr{P}_A$ be a collection of such partitions, i.e.

$$\mathscr{P}_{_A} = \left\{ \mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P} : \mathbf{I} = \{\mathbf{I}_l\}_{l=1}^k \text{ satisfies } n\hat{p}_{_{j \cdot l}} > 5; \text{ for all } j \in \{g, g'\} \text{ and } l = 1, \dots, k \right\}. \quad (6)$$

For the remainder of the paper, we shall refer to $\mathscr{P}_{_A}$ as a collection of *admissible partitions*. Note that $n\hat{p}_{_{j \cdot l}} > 5$ is the only restriction on the structure of $\mathbf{P}_n^{(k)}(\mathbf{I})$ in (2), therefore there are many admissible partitions $\mathbf{P}_n^{(k)}(\mathbf{I})$ that can be formed from the observed joint data $\{(d_{g \cdot i},\, d_{g' \cdot i})\}_{i=1}^n$. As $\mathscr{P}$ is finite, $\mathscr{P}_{_A}$ is also a finite set of partitions. In such a context, proving the uniform control of *type-I error* over $\mathscr{P}_{_A}$ of the statistics considered for assessing H_{0m} in (1) is important.

Now, let

$$\begin{aligned} u_{j \cdot i} &= [\mathbb{1}(d_{j \cdot i} \in \mathbf{I}_1),\ \dots,\ \mathbb{1}(d_{j \cdot i} \in \mathbf{I}_k)]', \quad p_{_{j \cdot i}} = \mathbb{E}(u_{j \cdot i}) \equiv [p_{_{j \cdot i1}}, \dots, p_{_{j \cdot ik}}]', \quad (7) \\ \text{and } \hat{p}_{_j} &:= [\hat{p}_{_{j \cdot 1}}, \dots, \hat{p}_{_{j \cdot k}}]' = \frac{1}{n} \sum_{i=1}^n u_{j \cdot i},\ j \in \{g, g'\}, \quad (8) \end{aligned}$$

where $p_{_{j \cdot il}}$ and $\hat{p}_{_{j \cdot l}}$ are given in (4). Each estimated vector of probabilities $\hat{p}_{_j}$ in (8) is a *sample average* of the realizations $u_{j \cdot i}$ from a k-dimensional multinomial random variable with vector of parameters $p_{_{j \cdot i}} = [p_{_{j \cdot i1}}, \dots, p_{_{j \cdot ik}}]'$. Let $\widehat{\Sigma}_j$ be an estimator of the covariance matrix of $u_{j \cdot i}$ given by

$$\widehat{\Sigma}_j = \begin{pmatrix} \hat{p}_{_{j \cdot 1}}(1 - \hat{p}_{_{j \cdot 1}}) & -\hat{p}_{_{j \cdot 1}}\hat{p}_{_{j \cdot 2}} & \dots & -\hat{p}_{_{j \cdot 1}}\hat{p}_{_{j \cdot k}} \\ -\hat{p}_{_{j \cdot 2}}\hat{p}_{_{j \cdot 1}} & \hat{p}_{_{j \cdot 2}}(1 - \hat{p}_{_{j \cdot 2}}) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ -\hat{p}_{_{j \cdot k}}\hat{p}_{_{j \cdot 1}} & -\hat{p}_{_{j \cdot k}}\hat{p}_{_{j \cdot 2}} & \dots & \hat{p}_{_{j \cdot k}}(1 - \hat{p}_{_{j \cdot k}}) \end{pmatrix} : \ j \in \{g, g'\}, (9)$$

and similarly, define

$$p_{_{gg' \cdot i, l\tilde{l}}} = \mathbb{P}(d_{g \cdot i} \in \mathbf{I}_l, d_{g' \cdot i} \in \mathbf{I}_{\tilde{l}}), \ \ \hat{p}_{_{gg' \cdot l\tilde{l}}} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(d_{g \cdot i} \in \mathbf{I}_l)\mathbb{1}(d_{g' \cdot i} \in \mathbf{I}_{\tilde{l}}), \quad (10)$$

and let $\widehat{\Sigma}_{gg'}$ be an estimator of the covariance matrix of the $(2k)$-dimensional vector of joint variables $(u'_{g\cdot i} \,:\, u'_{g'\cdot i})'$ given by

$$\widehat{\Sigma}_{gg'} \quad = \quad \begin{pmatrix} \hat{p}_{gg'\cdot 11} - \hat{p}_{g\cdot 1}\hat{p}_{g'\cdot 1} & \hat{p}_{gg'\cdot 12} - \hat{p}_{g\cdot 1}\hat{p}_{g'\cdot 2} & \cdots & \hat{p}_{gg'\cdot 1k} - \hat{p}_{g\cdot 1}\hat{p}_{g'\cdot k} \\ \hat{p}_{gg'\cdot 21} - \hat{p}_{g\cdot 2}\hat{p}_{g'\cdot 1} & \hat{p}_{gg'\cdot 22} - \hat{p}_{g\cdot 2}\hat{p}_{g'\cdot 2} & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \hat{p}_{gg'\cdot k1} - \hat{p}_{g\cdot k}\hat{p}_{g'\cdot 1} & \hat{p}_{gg'\cdot k2} - \hat{p}_{g\cdot k}\hat{p}_{g'\cdot 2} & \cdots & \hat{p}_{gg'\cdot kk} - \hat{p}_{g\cdot k}\hat{p}_{g'\cdot k} \end{pmatrix}. \tag{11}$$

Also, let $\hat{v}_m = \mathbf{T}^m(\hat{p}_g - \hat{p}_{g'})$ be the scaled vector of contrasts, where $\mathbf{T}$ is a $k \times k$ lower triangular matrix of ones, $\mathbf{T}^m$ denotes the matrix $\mathbf{T}$ to the m-th power, and define

$$\widehat{\Omega}_m = \mathbf{T}^m[\widehat{\Sigma}_g + \widehat{\Sigma}_{g'} - (\widehat{\Sigma}_{gg'} + \widehat{\Sigma}'_{gg'})]\mathbf{T}'^m := [\widehat{\omega}_{m\cdot l\tilde{l}}]_{1\leqslant l,\tilde{l}\leqslant k} \,. \tag{12}$$

Note that by construction, each of the $k \times k$ matrices $\widehat{\Sigma}_j, j \in \{g, g'\}$ in (9), $\widehat{\Sigma}_{gg'}$ in (11), and $\widehat{\Sigma}_g + \widehat{\Sigma}_{g'} - \widehat{\Sigma}_{gg'} - \widehat{\Sigma}'_{gg'}$ in (12) have rank $k-1$. Therefore, $\widehat{\Omega}_m$ in (12) also has rank $k-1$. The notation $\widehat{\Omega}_m^-$ thus refers to the *generalized inverse* of $\widehat{\Omega}_m$ hereinafter. From Dhrymes (1978, Proposition 3.5), there exists a diagonal matrix $\widehat{D}_{k-1}$ whose diagonal elements are the nonzero eigenvalues of $\widehat{\Omega}_m$ (in decreasing order of magnitude), and a $k \times (k-1)$ matrix $\widehat{P}_{k-1}$ whose columns are the (orthogonal) eigenvectors corresponding to the nonzero roots of $\widehat{\Omega}_m$, such that

$$\widehat{\Omega}_m^- = \widehat{P}_{k-1}\widehat{D}_{k-1}^{-1}\widehat{P}'_{k-1}. \tag{13}$$

We now make the following assumption on the joint sample $\{(d_{g\cdot i},\, d_{g'\cdot i})\}_{i=1}^n$.

Assumption 1. $\mathscr{D}_n := \{(d_{g\cdot i}, d_{g'\cdot i})\}_{i=1}^n$ *is a i.i.d. random sample across i drawn from the joint distribution of the degrees of networks g and g'.*

In the above assumption, possible dependence between the distribution of the degrees of the two networks is allowed. The i.i.d. sampling across the rows of the joint sample $\mathscr{D}_n$ preserves this dependence. In the case where g and g' are independent, one can draw two independent i.i.d. samples with different sizes: one from the population of network g, say $(d_{g\cdot i})_{i=1}^{n_g}$, and the second from the population of network g', say $(d_{g'\cdot i})_{i=1}^{n_{g'}}$. However, this case excludes interesting situations where the populations of the two networks overlap, as is usually the case in most empirical applications of social networks. In such contexts, while it is reasonable to assume that $(d_{g\cdot i}, d_{g'\cdot i})$ is independent of $(d_{g\cdot i'}, d_{g'\cdot i'})$ for $i \neq i'$, it is likely that $d_{g\cdot i}$ and $d_{g'\cdot i}$ will be correlated.

4. Test Statistics and asymptotic theory

We wish to first discuss how problem (1) can be recast in the more familiar language of hypotheses specified on vectors of contrast. Under the i.i.d. sampling across observations in Assumption 1, we have $p_{_{j\cdot il}} = p_{_{j\cdot l}}$ in (4) and $p_{_{j\cdot i}} = p_{_j}$ in (7) for all $j \in \{g, g'\}$, $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, k\}$. Therefore, it is straightforward to show that problem (1) can be equivalently formulated[14] as:

$$H_{0m} : v_m = 0 \text{ versus } H_{1m} : v_m \leqslant 0 \;\wedge\; H_{2m} : v_m \nleqslant 0 \text{ and } v_m \ngeqslant 0 \quad (14)$$

for any $m \in \mathbb{N}$, where $v_m = \mathbf{T}^{^m}(p_{_g} - p_{_{g'}})$ and $\mathbf{T}$ is given in (12). Since v_m is a $k \times 1$ scaled vector of contrasts, testing H_{0m} in (14) involves k multiple comparison procedures and there is a risk of size control related to

[14] See Anderson (1996) for a similar formulation.

a simultaneous testing of the significance of *pairwise contrasts*. To avoid size distortions, Richmond (1982) proposes to use the studentized maximum modulus (SMM) type statistic whose distribution is tabulated by Stoline and Ury (1979), and the statistic is employed by Beach and Richmond (1985) to construct confidence regions for Lorenz curve ordinates. In this paper, we combine the studentized maximum modulus statistic with an adjusted version of Pearson's (1900) statistic for assessing problem (14). Anderson (1996) employed a similar method in the context of income distributions but his methodology relies on the assumption that $(d_{g\cdot i})_{i=1}^{n}$ and $(d_{g'\cdot i})_{i=1}^{n}$ are independent, while ours is free of such a restriction.

To be more specific, suppose that Assumption 1 is satisfied. Hence, we have $\hat{p}_{_g} \xrightarrow{p} p_{_g}$ and $\hat{p}_{_{g'}} \xrightarrow{p} p_{_{g'}}$, so that the estimated contrast $\hat{v}_m = \mathbf{T}^{^m}(\hat{p}_{_g} - \hat{p}_{_{g'}}) \xrightarrow{p} v_m = \mathbf{T}^{^m}(p_{_g} - p_{_{g'}})$. If further H_{0m} holds, $v_m = 0$ and $\hat{v}_m$ will be close to zero for a large enough sample size. However, under H_{1m} or H_{2m}, neither v_m nor $\hat{v}_m$ will be close to zero. Therefore, one can detect whether H_{0m} is violated by looking at how far the estimated contrast $\hat{v}_m$ is from zero. Since the estimated contrast $\hat{v}_m$ will not be exactly zero under H_{0m} due to sampling error, a conventional way to proceed is to construct the test statistic from the distribution of $\hat{v}_m$. This approach is extensively discussed in Hausman (1978) and widely used in econometrics, especially in specification testing. Before we move on to the derivation of the statistics for H_{0m}, it is useful to establish the following convergence property for the estimated contrast of probabilities $\hat{p}_{_g} - \hat{p}_{_{g'}}$, as well as its scaled variant $\hat{v}_m = \mathbf{T}^{^m}(\hat{p}_{_g} - \hat{p}_{_{g'}})$.

Lemma 1. *Suppose that Assumption 1 holds. For any admissible partition*

$\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_{A}$, *we have:*

$$\sqrt{n}[(\hat{p}_{g} - \hat{p}_{g'}) - (p_{g} - p_{g'})] \overset{d}{\to} N\Big(0,\ \Sigma_g + \Sigma_{g'} - \Sigma_{gg'} - \Sigma'_{gg'}\Big), \tag{15}$$

$$\textit{and}\ \sqrt{n}(\hat{v}_m - v_m) \overset{d}{\to} N(0,\ \Omega_m), \tag{16}$$

where $\Sigma_j = \underset{n\to\infty}{p\lim}(\widehat{\Sigma}_j)$, $j \in \{g, g'\}$, $\Sigma_{gg'} = \underset{n\to\infty}{p\lim}(\widehat{\Sigma}_{gg'})$, $\Omega_m = \mathbf{T}^{m}(\Sigma_g + \Sigma_{g'} - \Sigma_{gg'} - \Sigma'_{gg'})\mathbf{T}'^{m}$, $\widehat{\Sigma}_j$ *and* $\widehat{\Sigma}_{gg'}$ *are defined in* (9) - (11).

Lemma 1 follows by the multivariate central limit theorem (MVCLT) property and the proof is presented in the appendix. It states that the estimated contrast $(\hat{p}_{g} - \hat{p}_{g'})$ and its scaled variant $\hat{v}_m$ are *root-n* consistent and asymptotically normal. Anderson (1996) assumes that $\Sigma_{gg'} = 0$, so we have $\Omega_m = \mathbf{T}^{m}(\Sigma_g + \Sigma_{g'})\mathbf{T}'^{m}$ in his setup. In the context of correlated samples ($\Sigma_{gg'} \neq 0$), a correction to Anderson's (1996) statistics is necessary to avoid size distortions, and the term $-(\Sigma_{gg'} + \Sigma'_{gg'})$ on the rhs of (16) is the adjustment needed.[15] In the appendix (see Lemma 4), we show that Ω_m can be consistently estimated by $\widehat{\Omega}_m = \mathbf{T}^{m}(\widehat{\Sigma}_g + \widehat{\Sigma}_{g'} - \widehat{\Sigma}_{gg'} - \widehat{\Sigma}'_{gg'})\mathbf{T}'^{m}$, where $\widehat{\Sigma}_j, j \in \{g, g'\}$ and $\widehat{\Sigma}_{gg'}$ are given in (9) - (11). Observe that $\widehat{\Sigma}_{gg'}$ is built from the contingency table obtained from the partition $\mathbf{P}_n^{(k)}(\mathbf{I})$ (thus from the distribution of the joint sample), while $\widehat{\Sigma}_j$ only exploits the information from the marginal distribution of the sample of network $j \in \{g, g'\}$. The main conclusion here is that even though the cdfs (hence the pdfs) of the two networks are equal under H_{0m}, constructing the Pearson- or SMM-type statistics solely based on them, as is usually done in the literature on inequality and poverty measures, is not always the best way to go because it does not account for the correlation structure between networks.

[15] Our investigation through a Monte Carlo experiment shows that failing to adjust Anderson's (1996) statistics yields overly size distorted tests when the two samples $(d_{g\cdot i})_{i=1}^{n}$ and $(d_{g'\cdot i})_{i=1}^{n}$ are correlated. In order to shorten the exposition, this exercise is omitted from this paper but it is available upon request.

We now focus on the derivation of the test statistics for H_{0m}.

4.1. Test Statistics and Decision Rule

Following Anderson (1996), we consider two statistics based on the estimated vector of contrasts $\hat{v}_m$ for assessing H_{0m}:

$$\begin{aligned} \mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) &= n\hat{v}_m'\widehat{\Omega}_m^{-}\hat{v}_m = n\hat{v}_m'\widehat{P}_{k-1}\widehat{D}_{k-1}^{-1}\widehat{P}_{k-1}'\hat{v}_m, \\ \mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) &= \max_{1\leqslant l\leqslant k-1}\left(|\sqrt{n}\widehat{Z}_{ml}|\right), \end{aligned} \tag{17}$$

where $\widehat{Z}_{ml}$ is the lth component of $\widehat{D}_{k-1}^{-1/2}\widehat{P}_{k-1}'\hat{v}_m$, $\widehat{D}_{k-1}$ and $\widehat{P}_{k-1}$ are given in (13).[16] $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ in (17) is a *Pearson*-type statistic expressed as a quadratic form in $\hat{v}_m$. It differs from that in Anderson (1996) not only through the correction of the covariance matrix $\widehat{\Omega}_m$, but also its direct dependence on $\mathbf{P}_n^{(k)}(\mathbf{I})$. The dependence on $\mathbf{P}_n^{(k)}(\mathbf{I})$ underscores the importance of controlling the size of the resulting test uniformly over the collection of admissible partitions $\mathscr{P}_A$. Uniformity over $\mathscr{P}_A$ is crucial for the asymptotic results to give a good approximation of the empirical size of the tests to the nominal level. $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ is a generalization of the SMM statistic in Stoline and Ury (1979). Besides its dependence on $\mathbf{P}_n^{(k)}(\mathbf{I})$, the expression of $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ in (17) is conceptually different from those in Stoline and Ury (1979), Beach and Richmond (1985), and Anderson (1996). For example, Beach and Richmond (1985) and Anderson (1996) defined these statistics as $\max\limits_{1\leqslant l\leqslant k}\left(\widehat{\omega}_{m\cdot ll}^{-1/2}|\sqrt{n}\hat{v}_{ml}|\right)$, where $\hat{v}_{ml}$ is the lth component of $\hat{v}_m$ and $\widehat{\omega}_{m\cdot ll}$ is the $(l,l)^{th}$ element of $\widehat{\Omega}_m$. Since $\hat{v}_{ml}$ and $\widehat{\omega}_{m\cdot ll}$ are not independent by con-

[16] The Wald statistic in equation (17) may also be written as $\mathbf{W}_m\left(\mathbf{P}_n^k(\mathbf{I})\right) = n(\hat{p}_g - \hat{p}_{g'})'\left[\hat{\Sigma}_g + \hat{\Sigma}_{g'} - \hat{\Sigma}_{gg'} - \hat{\Sigma}_{gg'}'\right]^{-}(\hat{p}_g - \hat{p}_{g'})$. That is, it is invariant to the choice of m.

struction,[17] $\max\limits_{1\leqslant l\leqslant k}\left(\widehat{\omega}_{m\cdot ll}^{-1/2}|\sqrt{n}\hat{v}_{ml}|\right)$ does not follow a SMM distribution under H_{0m}. By contrast, the expression of $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ in (17) converges to a SMM distribution with parameter $k-1$ and infinite degrees of freedom under H_{0m} and Assumption 1 (see Lemma 2). This is because we have adjusted this statistic as the maximum of the absolute values of $k-1$ non-redundant linear combinations of the components of $\sqrt{n}\hat{v}_m$, where the weights are the elements of the $(k-1)\times k$ matrix $\widehat{D}_{k-1}^{-1/2}\widehat{P}_{k-1}'$, while $\max\limits_{1\leqslant l\leqslant k}\left(\widehat{\omega}_{m\cdot ll}^{-1/2}|\sqrt{n}\hat{v}_{ml}|\right)$ is obtained as the maximum of the absolute value of the k component of the scaled vector $[\mathbf{diag}(\widehat{P}_{k-1}\widehat{D}_{k-1}^{-1}\widehat{P}_{k-1}')]^{1/2}\sqrt{n}\hat{v}_m$. Moreover, one of the fundamental differences between the two statistics in (17) is that $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ does not depend on either $\mathbf{T}$ or m (order of dominance tested),[18] while $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ depends on both.

Since $\hat{v}_m \xrightarrow{p} v_m$ under Assumption 1, it is clear from (14) that $F_g \succ_m F_{g'}$ if all components of $\hat{v}_m$ are less or equal to zero, with a strict inequality at least for one. Hence, the statistic $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$, which is a quadratic form in $\hat{v}_m$, if not combined with $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$, tests the equality between the cumulative distributions F_g and $F_{g'}$ and a rejection does not necessary entail *stochastic dominance*. Meanwhile, a rejection using the statistic $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ implies *stochastic dominance*. Furthermore, the test with $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ asymptotically controls the "familywise" rate of type I error in multiple comparison procedures (e.g., Richmond (1982) and Beach and

[17] The upper α-points of the distribution of the SMM statistic, $\max\limits_{1\leqslant l\leqslant k}\left(\widehat{\omega}_{m\cdot ll}^{-1/2}|\sqrt{n}\hat{v}_{ml}|\right)$, in Stoline and Ury (1979, Tables 1-3) are provided under the assumption that $\hat{v}_{ml}$ is independent of $\widehat{\omega}_{m\cdot ll}$. However, the partitioning into classes does not preserve this independence assumption.

[18] As $\mathbf{T}$ is invertible, $\mathbf{T}^m$ is also invertible for all $m\in\mathbb{N}$ so that $\hat{v}_m'\widehat{\Omega}_m^{-}\hat{v}_m = \hat{v}'\mathbf{T}^{m'}\mathbf{T}^{-m'}\widehat{\Omega}^{-}\mathbf{T}^{-m}\mathbf{T}^m\hat{v} = \hat{v}'\widehat{\Omega}^{-}\hat{v}$, i.e., $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ does not depend on either $\mathbf{T}$ nor m.

Richmond (1985)). A combination of the two statistics informs us on whether '*stochastic dominance*' holds or not, once equality between the two distributions is rejected. Formally, as long as the two statistics are combined, one of the following three levels of decision can be reached given any admissible partition $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$:

1. if $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant c_k(\alpha)$, retain H_{0m};
2. if $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > c_k(\alpha)$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > s_k(\alpha)$, retain H_{1m};
3. if $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > c_k(\alpha)$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant s_k(\alpha)$, retain H_{2m},

where for some $\alpha \in (0,\ 1)$, the cut-off points $c_k(\alpha)$ and $s_k(\alpha)$ are determined such that $\mathbb{P}[\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > c_k(\alpha)] \to \alpha$ and $\mathbb{P}[\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > s_k(\alpha)] \to \alpha$ under H_{0m}, as $n \to \infty$ (at least). Tests based on the two statistics are not equally powerful against both alternatives H_{1m} and H_{2m}, especially in small samples. Indeed, in the case where one cumulative distribution is completely below the other, both tests have good power. However, if the cumulative distributions cross, the test with $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ is more powerful than those with $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$. This is because $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ is a quadratic form in $\sqrt{n}\widehat{Z}_m = \widehat{D}_{k-1}^{-1/2}\widehat{P}_{k-1}'\sqrt{n}\widehat{v}_m$ while $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ is the absolute value of the maximal component of $\sqrt{n}\widehat{Z}_m \in \mathbb{R}^{k-1}$. Furthermore, from the functional forms of $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ in (17), a non-rejection by the test with $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ entails a non-rejection of those with $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$, as long as the tests are performed at the same nominal level. Thus, retaining H_{0m} when the test with $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ fails to reject it asymptotically controls the "familywise" rate of type I error. Hence, Bonferroni-type size correction for multiple comparison hypotheses is not warranted in large samples. To enhance the small-sample performance of the test, we propose a bootstrap method that is easy to implement from the observed data (see Section 4.3). But before we move on to the bootstrap results, it is informative to study the asymptotic properties of the standard tests first.

4.2. Asymptotic Properties of the tests

In this section, we characterize the large-sample properties (size and power) of the above tests of stochastic dominance. To do this, we first study the asymptotic behavior of $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ under both the null hypothesis (H_{0m}) and the alternative hypotheses (H_{1m} and H_{2m}). Lemma 2 presents the results.

Lemma 2. *Let $\mathbf{P}_n^{(k)}(\mathbf{I})$ be any admissible partition in $\mathscr{P}_A$. Under Assumption 1, the following convergence results hold as n goes to infinity:*

(*a*) *if H_{0m} is satisfied, we have*

$$\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{d} \chi^2(k-1), \quad \mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{d} \max_{1 \leqslant l \leqslant k-1} |\mathscr{Z}_l| \sim SMM(k-1, \infty),$$

(*b*) *if H_{1m} or H_{2m} is satisfied, we have*

$$\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{p} +\infty, \quad \mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{p} +\infty,$$

where $\mathscr{Z}_l \overset{i.i.d.}{\sim} N(0,1)$ for all $l = 1, 2, \ldots k-1$ and $SMM(k-1, \infty)$ is the studentized maximum modulus distribution with parameter $k-1$ and infinite degrees of freedom.

Lemma 2 - (a) shows that for any admissible partition $\mathbf{P}_n^{(k)}(\mathbf{I})$ in $\mathscr{P}_A$, the asymptotic distributions under H_{0m} of both statistics are nuisance parameters free. The statistic $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ has the standard χ^2 asymptotic distribution, while that of $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ is non-standard but its critical values are tabulated in Stoline and Ury (1979). Lemma 2 - (b) indicates that the statistics diverge under H_{1m} or H_{2m} for any admissible partition $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$. We can now establish the following results on the uniform control of the size over $\mathscr{P}_A$ as well as test consistency for any partition $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$.

Theorem 1. *Suppose that Assumption 1 is satisfied and let $\alpha \in (0,1)$. As the sample size n goes to infinity, the following convergence results holds*:

(*a*) *if H_{0m} is satisfied, then we have*

$$\limsup_{n\to\infty} \sup_{\mathscr{P}_A} \mathbb{P}[\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > \chi^2_{k-1}(\alpha)] = \alpha, \ \limsup_{n\to\infty} \sup_{\mathscr{P}_A} \mathbb{P}[\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > z_{k-1}(\alpha)] = \alpha;$$

(*b*) *if H_{1m} or H_{2m} is satisfied, then we have*

$$\lim_{n\to\infty} \mathbb{P}[\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > \chi^2_{k-1}(\alpha)] = 1, \quad \lim_{n\to\infty} \mathbb{P}[\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > z_{k-1}(\alpha)] = 1 \ \forall \ \mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A,$$

where $\chi^2_{k-1}(\alpha)$ and $z_{k-1}(\alpha)$ are the $(1-\alpha)^{th}$ quantiles of a $\chi^2(k-1)$-distributed and a $SMM(k-1,\infty)$-distributed random variables, respectively.

Theorem 1-(a) shows that tests based on both $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ have correct size uniformly over $\mathscr{P}_A$. Therefore, the asymptotic χ^2 and SMM critical values provide good approximations of the empirical critical values of $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ if n is large. Theorem 1-(b) indicates that both tests are consistent under H_{1m} or H_{2m} for any admissible partition $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$. However, the finite-sample size and power of the tests depend on the choice of $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$, and may not be as good as their asymptotic properties. To address this issue, we propose a bootstrap method to enhance the finite-sample properties of the tests. Section 4.3 presents the details.

4.3. Bootstrap Tests

In this section, we study the validity of the bootstrap for the statistics $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$. The usual intuition for the bootstrap requires that the empirical distribution, from which the bootstrap sample is drawn, be close to the distribution of the data under the null hypothesis. In our context, the empirical distribution used in the bootstrap sampling is the empirical distribution of the joint sample $\mathscr{D}_n = \{(d_{g\cdot i},\, d_{g'\cdot i})\}_{i=1}^{n}$. To be more specific, the bootstrap pseudo-samples and statistics, as well as the decision rule are obtained following the above steps.

1. From the observed joint sample $\mathscr{D}_n = \{(d_{g\cdot i},\, d_{g'\cdot i})\}_{i=1}^{n}$, obtain a partition $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_{A}$ and compute the realizations of the statistics $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$.

2. For each bootstrap sample $b = 1, \ldots, M_b$, generate the data $\mathscr{D}_n^{*} = \{(d_{g\cdot i}^{*},\, d_{g'\cdot i}^{*})\}_{i=1}^{n}$, where $(d_{g\cdot i}^{*}, d_{g'\cdot i}^{*})$ are drawn independently from the empirical distribution of the joint sample $\mathscr{D}_n$. From the re-sampled data and the partition $\mathbf{P}_n^{(k)}(\mathbf{I})$, compute the realizations of the bootstrap statistics $\mathbf{W}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})),\ \mathbf{S}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) :\ b = 1, \ldots, M_b$:

$$\mathbf{W}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) \;=\; n\tilde{v}_m^{*\prime}\widehat{\Omega}_m^{*-}\tilde{v}_m^{*},\ \mathbf{S}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) = \max_{1\leqslant l\leqslant k-1}\Big(|\sqrt{n}\widetilde{Z}_{ml}^{*}|\Big) \tag{18}$$

   where $\tilde{v}_m^{*} = \hat{v}_m^{*} - \hat{v}_m$, $\widetilde{Z}_{ml}^{*} = \widehat{Z}_{ml}^{*} - \widehat{Z}_{ml}$; and $\widehat{\Omega}_m^{*-}$, $\hat{v}_m^{*}$, $\widehat{Z}_{ml}^{*}$ are the bootstrap analogues of $\widehat{\Omega}_m^{-}$, $\hat{v}_m$, $\widehat{Z}_{ml}$, respectively.

3. The decision rule of the bootstrap test is as follows:

   (a) if $\frac{1}{M_b}\sum_{b=1}^{M_b}\mathbb{1}\big[\mathbf{W}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) > \mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))\big] \geqslant \alpha$ where $\mathbb{1}[C] = 1$ if condition C holds and $\mathbb{1}[C] = 0$ otherwise, retain H_{0m};

(b) if $\frac{1}{M_b}\sum_{b=1}^{M_b}\mathbb{1}\big[\mathbf{W}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) > \mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))\big] < \alpha \wedge \frac{1}{M_b}\sum_{b=1}^{M_b}\mathbb{1}\big[\mathbf{S}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) > \mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))\big] < \alpha$, retain H_{1m};

(c) if $\frac{1}{M_b}\sum_{b=1}^{M_b}\mathbb{1}\big[\mathbf{W}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) > \mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))\big] < \alpha \wedge \frac{1}{M_b}\sum_{b=1}^{M_b}\mathbb{1}\big[\mathbf{S}_m^{*(b)}(\mathbf{P}_n^{(k)}(\mathbf{I})) > \mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))\big] \geqslant \alpha$, retain H_{2m}.

The bootstrap statistics in (18) are expressed in terms of $\tilde{v}_m^* = \hat{v}_m^* - \hat{v}_m$, rather than $\hat{v}_m^*$. This re-centering is important for the validity of the bootstrap as the expectation of $\hat{v}_m^*$ under the bootstrap data generating process is $\hat{v}_m$, which is not necessarily zero under H_{0m}. The importance of re-centering has extensively been discussed in the bootstrap literature (e.g., Hall and Horowitz (1996), Hahn (1996), Andrews (2002), Brown and Newey (2002), Inoue and Shintani (2006)).

In the remainder of the paper, the probability under the empirical distribution function of the joint sample $\mathscr{D}_n^*$ conditional on the observed data $\mathscr{D}_n$ is denoted by $\mathbb{P}^*[\cdot]$, and $\mathbb{E}^*[\cdot]$ is its corresponding expectation operator. Lemma 3 characterises the asymptotic behavior of the bootstrap statistics of stochastic dominance.

Lemma 3. *Let $\mathbf{P}_n^{(k)}(\mathbf{I})$ be any admissible partition in $\mathscr{P}_A$. Under Assumption 1, the following convergence results hold as n goes to infinity:*

(a) if H_{0m} is satisfied, then we have

$$\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{d} \chi^2(k-1) \ a.s., \ \mathbf{S}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{d} \max_{1\leqslant l\leqslant k-1}|\mathscr{Z}_l| \sim SMM(k-1,\infty) \ a.s$$

(b) if H_{1m} or H_{2m} is satisfied, then we have

$$\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \mid \mathscr{D}_n \ \xrightarrow{p} \ +\infty \ a.s. \ \mathbf{S}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{p} +\infty \ a.s.,$$

where $\mathscr{Z}_l$ and $SMM(k-1,\infty)$ are defined in Lemma 2.

Lemma 3 shows that the bootstrap provides a first-order approximation of the null limiting distributions of the statistics $\mathbf{W}^*_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $\mathbf{S}^*_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$, and is further consistent under the alternative hypotheses H_{1m} and H_{2m}. These results hold irrespective of which partition $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$ is used in the computation of the statistics. We can prove the following theorem on the consistency of the bootstrap tests.

Theorem 2. *Let $\mathbf{P}_n^{(k)}(\mathbf{I})$ be any admissible partition in $\mathscr{P}_A$, and suppose that Assumption 1 is satisfied. Then, the following convergence results hold as n goes to infinity, whether H_{0m} holds or not*:

$$\sup_{w\in\mathbb{R}}\left|\mathbb{P}^*\big(\mathbf{W}^*_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w\big) - \mathbb{P}\big(\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w\big)\right| \;\to\; 0 \textit{ in probability } \mathbb{P},$$

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}^*\big(\mathbf{S}^*_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant z\big) - \mathbb{P}(\mathbf{S}_m\big(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant z\big)\right| \;\to\; 0 \textit{ in probability } \mathbb{P}.$$

We now study the finite-sample performance (size and power) of both the asymptotic and bootstrap tests of stochastic dominance through a Monte Carlo experiment.

5. Monte Carlo Experiment

In this section, we use simulation to examine the finite-sample size and power performance of both the asymptotic and bootstrap tests of stochastic dominance. To shorten the exposition, we only present the results for $m = 2$ in (1). So, the null hypothesis (H_{02}) tests the equality between the two networks' distributions against second-order stochastic dominance (H_{12}), or no second-order stochastic dominance (H_{22}). The data generating process (DGP) covers the most common distributions that are used in applied work to model the degrees of networks. Precisely, the two DGPs are specified as follows.

(I). $(d_{g\cdot i}, d_{g'\cdot i})', i = 1, \ldots, n$, are drawn i.i.d. across i from a *bivariate Poisson distribution* with mean $(10, \lambda)'$ and correlation ρ. In this setup, the null hypothesis that the cdfs of $\left(d_{g\cdot i}\right)_{i=1}^{n}$ and $\left(d_{g'\cdot i}\right)_{i=1}^{n}$ are equal can be expressed as $\lambda = 10$. So, $\lambda \neq 10$ describes either H_{12} or H_{22}.

(II). $(d_{g\cdot i}, d_{g'\cdot i})', i = 1, \ldots, n$, are drawn i.i.d. across i from a *bivariate Scale-free distribution*[19] with parameters $(2.5, \gamma)'$ and correlation ρ. As in design **(I)**, the cdfs of $\left(d_{g\cdot i}\right)_{i=1}^{n}$ and $\left(d_{g'\cdot i}\right)_{i=1}^{n}$ are equal for a given ρ if and only if $\gamma = 2.5$. So, the values of $\gamma \neq 2.5$ characterize a violation of the null hypothesis.

In both setups, we vary ρ (correlation between the two samples) in $\{-0.9, -0.5, 0, 0.5, 0.9\}$, but the results do not change qualitatively with alternative choices of ρ. In all cases, the joint sample is generated using the algorithm provided by Macke et al. (2009) and Bethge and Berens (2007). As noted in Figure 2, the support of the Poisson distribution with $\lambda = 10$ is in the range 1-20, while that of the Scale-free distribution with $\gamma = 2.5$ is in the range 1-9. Hence, any admissible partition may take these ranges into account. In order to shorten the exposition, we consider two partitions for each setup. In design **(I)**, the two partitions are $k = 4$ and $k = 8$, while they are $k = 3$ and $k = 4$ in design **(II)**. Specifically, $\mathbf{P}_n^{(4)}(\mathbf{I}) := \{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4\} = \{\{1, \ldots, 9\}, \{10\}, \{11\}, \{12+\}\}$ and $\mathbf{P}_n^{(8)}(\mathbf{I}) := \{\mathbf{I}_1, \ldots, \mathbf{I}_8\} = \{\{1, \ldots, 7\}, \{8\}, \{9\}, \{10\}, \{11\}, \{12\}, \{13\}, \{14+\}\}$ in design **(I)**, and in design **(II)** we have $\mathbf{P}_n^{(3)}(\mathbf{I}) := \{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3\} = \{\{1\}, \{2\}, \{3+\}\}$ and $\mathbf{P}_n^{(4)}(\mathbf{I}) :=$

[19] Note that the probability density function of a random variable D that follows a Scale-free distribution is given by $P(d) = d^{-\gamma}[\zeta(\gamma)]^{-1}, d \in \mathbb{N}$, where $\zeta(\gamma) = \sum_{d=1}^{+\infty} \frac{1}{d^{\gamma}}$ denotes the Riemann zeta function.

$\{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4\} = \{\{1\}, \{2\}, \{3\}, \{4+\}\}$. All these partitions belong to $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_{_A}$, and are thus admissible.

For the purpose of clarity and readability, we separate the analysis on the size from that on the power.

5.1. Size Properties

In this section, we analyze the empirical rejection frequencies of both the asymptotic and bootstrap tests of stochastic dominance for various sample sizes: $n \in \{100, 200, 500\}$. In each design and for each partition $\mathbf{P}_n^{(k)}(\mathbf{I})$ specified above, the statistics $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$, $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$, $\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$, and $\mathbf{S}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ are constructed as outlined in Sections 3.2, 4.1 & 4.3. The nominal level for both the asymptotic and bootstrap tests is set at $\alpha = 5\%$ and the empirical rejection frequencies are computed with $M = 10,000$ replications. The bootstrap critical values are approximated using $M_b = 199$ pseudo samples of size n. For the asymptotic tests, we use the $(1 - \alpha)^{th}$ quantiles of a $\chi^2(k-1)$ - distributed random variable for $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and a $SMM(k-1, \infty)$ - distributed random variable for $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$.

Table 3 presents the results of the two designs. The first column contains the partitions $\mathbf{P}_n^{(k)}(\mathbf{I})$, and the second shows both the asymptotic and bootstrap statistics. The other columns present, for each value of network endogeneity (ρ) and sample size n, the empirical rejection frequencies of the tests at the 5% nominal level.

First, in design **(I)** (Poisson distribution), the asymptotic tests are slightly size distorted for $n \in \{100, 200\}$. Their maximal size rejection frequencies is around 8.7% [for $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$] and 7.2% [for $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$] with the partition $\mathbf{P}_n^{(8)}(\mathbf{I})$, but they decrease with the partition $\mathbf{P}_n^{(4)}(\mathbf{I})$ (around

6.5% and 6.2% respectively). Meanwhile, their bootstrap counterparts have rejections close to the 5% nominal level in most cases for both partitions, even with $n = 100$. However, the bootstrap tests tend to under reject when $n = 100$ and $\rho = 0.9$, but this phenomenon disappears as the sample size increases. On top of its overall good performance in small samples, our results also suggest that the bootstrap tests are less sensitive to partitioning into classes than the asymptotic tests. Also, our results are consistent across all values of networks' endogeneity ρ.

Second, in design **(II)** (Scale-free distribution), both the asymptotic and bootstrap tests perform quite well irrespective of the partition used and network endogeneity ρ. However, the bootstrap tests tend to be conservative when $\rho = 0.9$ and $n \in \{100, 200\}$ while the empirical rejection frequencies of the asymptotic tests are consistently around the 5% nominal level for all sample sizes. Again, the under-rejections of the bootstrap tests observed when $\rho = 0.9$ and $n \in \{100, 200\}$ disappear as the sample size increases, as shown in the column $\rho = 0.9$ and $n = 500$ in the bottom part of the table.

5.2. Power Properties

We now study the empirical rejections of the various tests under the alternative hypothesis (power). For simplicity, we only present the power analysis for $n \in \{100, 500\}$ and $\rho \in \{0, 0.5, 0.9\}$. In design **(I)** (Poisson distribution), the power analysis is conducted in the direction of λ, where $\lambda = 10$ indicates the empirical size and $\lambda \neq 10$ indicates the empirical power. Similarly, the power analysis is conducted in the direction of γ in design **(II)** (Scale-free distribution): here $\gamma = 2.5$ indicates the empirical size, and $\gamma \neq 2.5$ characterizes the empirical power at γ.

Table 3: Empirical size of the asymptotic and bootstrap tests at 5%

(I): *Poisson distribution*																
		ρ =-0.9			ρ =-0.5			ρ =0			ρ =0.5			ρ =0.9		
$\mathbf{P}_n^{(k)}(\mathbf{I})$	n →	100	200	500	100	200	500	100	200	500	100	200	500	100	200	500
$\mathbf{P}_n^{(4)}(\mathbf{I})$	$\mathbf{W}_m$	5.7	5.1	5.4	6.0	5.8	5.0	6.2	6.0	5.0	6.5	5.6	5.0	6.0	5.5	5.0
	$\mathbf{S}_m$	5.6	5.3	5.1	5.8	5.8	5.0	5.8	5.3	5.1	6.2	5.4	4.7	5.8	5.2	5.2
	$\mathbf{W}^*_m$	4.5	4.7	5.0	4.7	5.2	4.9	4.7	5.2	4.8	5.1	4.8	4.7	3.6	4.8	4.8
	$\mathbf{S}^*_m$	4.4	4.8	5.0	4.6	5.1	4.9	4.8	4.7	4.9	4.9	4.7	4.7	3.6	4.6	4.8
$\mathbf{P}_n^{(8)}(\mathbf{I})$	$\mathbf{W}_m$	7.8	6.6	5.9	7.8	6.5	5.6	8.1	6.5	5.9	8.7	6.6	5.7	7.4	6.1	5.3
	$\mathbf{S}_m$	6.8	6.1	5.5	7.2	5.9	5.2	7.1	6.3	5.9	7.1	5.9	5.4	6.3	5.7	5.4
	$\mathbf{W}^*_m$	4.2	4.9	5.1	4.3	4.7	4.9	4.3	4.9	5.1	4.7	4.9	5.1	2.3	4.1	4.6
	$\mathbf{S}^*_m$	4.1	5.1	5.1	4.7	4.9	4.8	4.4	5.3	5.3	4.2	4.8	5.2	2.1	4.2	5.0
(II): *Scale-free distribution*																
		ρ =-0.9			ρ =-0.5			ρ =0			ρ =0.5			ρ =0.9		
$\mathbf{P}_n^{(k)}(\mathbf{I})$	n →	100	200	500	100	200	500	100	200	500	100	200	500	100	200	500
$\mathbf{P}_n^{(3)}(\mathbf{I})$	$\mathbf{W}_m$	5.5	5.2	5.2	6.0	5.6	5.0	5.6	5.2	5.3	5.5	5.4	5.4	4.3	4.9	5.1
	$\mathbf{S}_m$	5.4	5.1	5.3	5.8	5.5	4.8	5.6	5.2	5.0	5.7	5.4	5.2	4.8	5.1	5.2
	$\mathbf{W}^*_m$	4.7	4.8	5.2	5.2	5.2	4.9	5.0	4.9	5.1	4.3	4.9	5.4	1.6	3.5	4.9
	$\mathbf{S}^*_m$	4.6	4.7	5.2	5.0	5.2	4.8	5.0	4.9	4.6	4.5	4.9	5.1	1.2	3.7	5.0
$\mathbf{P}_n^{(4)}(\mathbf{I})$	$\mathbf{W}_m$	5.8	5.2	5.4	6.0	5.7	5.2	5.9	5.4	5.2	5.4	5.4	5.1	3.9	4.4	5.0
	$\mathbf{S}_m$	5.2	5.1	5.3	5.8	5.4	4.9	5.6	5.3	4.8	5.5	5.4	5.1	4.2	4.8	5.2
	$\mathbf{W}^*_m$	4.0	4.6	5.2	4.0	5.0	4.9	3.9	4.8	4.8	2.9	4.7	4.8	1.0	2.0	4.6
	$\mathbf{S}^*_m$	3.7	4.6	5.1	4.1	4.8	4.8	3.8	4.6	4.5	2.7	4.8	5.0	0.4	2.1	4.7

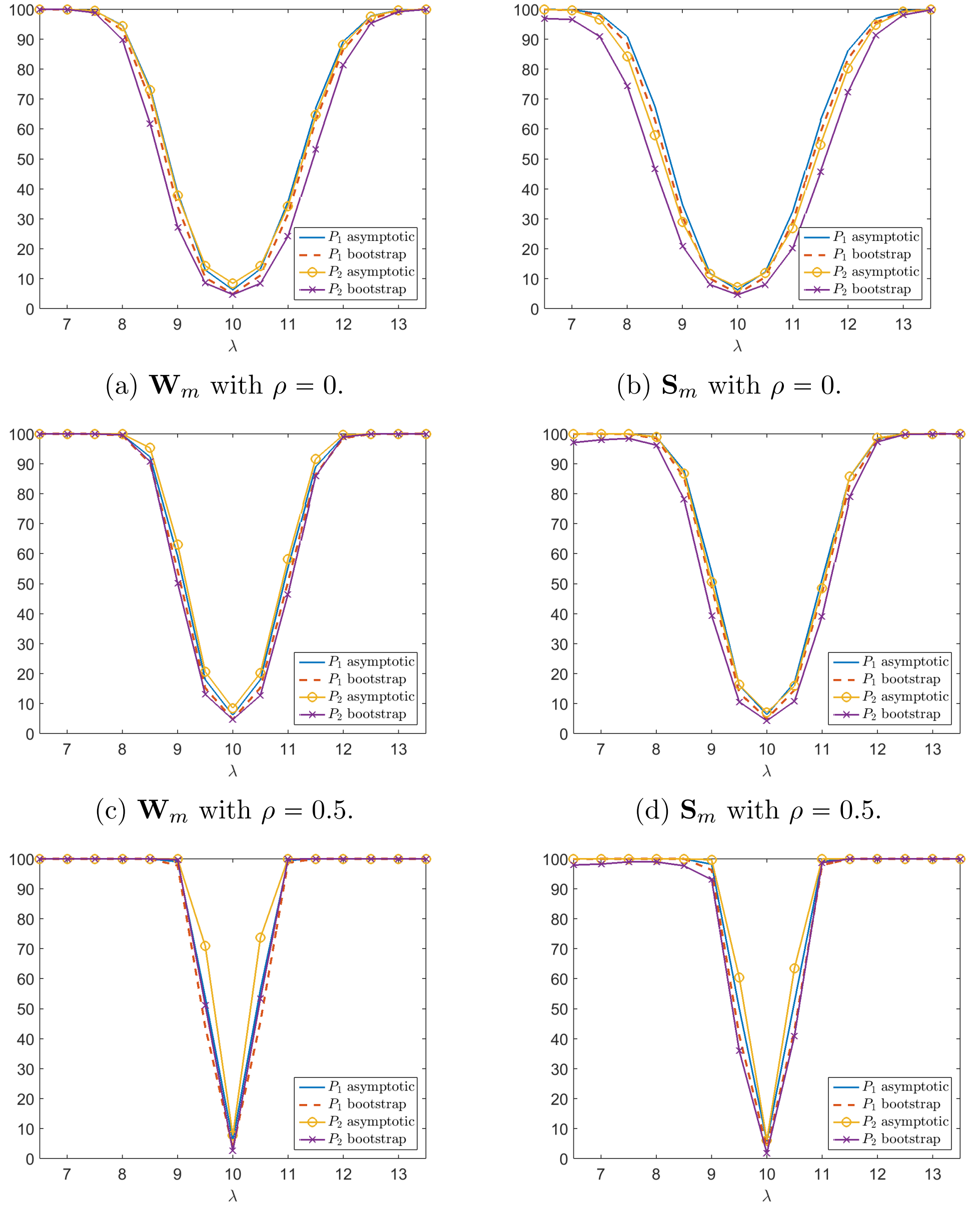
Figures 3 - 4 show the power curves of both the asymptotic and bootstrap tests in the two partitions for design **(I)**, while Figures 5 - 6 present similar graphs for design **(II)** (Scale-free distribution). In each figure, the sub-figures (a), (c), and (e) contain the power curves of $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and its bootstrap version, while the the sub-figures (b), (d), and (f) display the power curves of $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and its bootstrap version. Each sub-figure corresponds to a value of networks' endogeneity $\rho \in \{0, 0.5, 0.9\}$.

First, when $n = 500$ and for both designs, the asymptotic and the bootstrap tests perform similarly, irrespective of the value of ρ and the partition used (see Figure 4 and Figure 6). While the empirical power of all tests converges to 100% for large values of λ (Figure 4) and γ (Figure 6), the convergence is much lower in design **(II)** (Scale-free distribution) than in design **(I)** (Poisson distribution). This reflects the low speed of convergence in the approximation of a multinomial distribution to a multivariate normal distribution (see Lemma 1) when the original sample $\mathscr{D}_n$ is drawn from a Scale-free distribution than when it is drawn from a Poisson distribution. Although from the theory, both the asymptotic and bootstrap tests of stochastic dominance are consistent, knowing that the empirical power of tests approaches 1 with a sample size of $n = 500$ is an interesting result.

Second, when the sample size is relatively small (here $n = 100$), substantial differences between asymptotic and bootstrap tests appear. First, both the asymptotic and bootstrap tests exhibit more power in design **(I)** (Poisson distribution) than in design **(II)** (Scale-free distribution). For example, for independent networks ($\rho = 0$) or low correlated networks ($\rho = 0.5$), the empirical power is low for both the asymptotic and bootstrap tests in design **(II)** (see sub-figures (a), (b), (c) and (d) in Figure 5), while all tests exhibit more power in design **(I)** (see see sub-figures (a), (b), (c) and (d) in Figure 3). Second, within partitions, the asymptotic and bootstrap tests perform more similarly in design **(I)** than in design **(II)**. The slightly higher power of the asymptotic tests in Figure 3, especially for $\rho \in \{0, 0.5\}$ in partition $\mathbf{P}_n^{(8)}$, is due to their inability to control for the type-I error (see Table 3). Looking at the power of the bootstrap tests, partition $\mathbf{P}_n^{(4)}$ has a small edge over partition $\mathbf{P}_n^{(8)}$, especially for $\rho \in \{0, 0.5\}$. Mann and Wald (1942) and Williams (1950) recommended to allocate the same expected number in each cell, whilst maintaining a threshold of above 5 in order to optimize test power. Although both partitions $\mathbf{P}_n^{(4)}$ and $\mathbf{P}_n^{(8)}$ are admissible (in the sense that a threshold of above 5 is maintained in each cell), $\mathbf{P}_n^{(4)}$ is closer to Mann and Wald's (1942) and Williams's (1950) recommendation than $\mathbf{P}_n^{(8)}$ when it comes to allocate the same expected number in each cell. Note that the power gain from using $\mathbf{P}_n^{(4)}$ over $\mathbf{P}_n^{(8)}$ decreases as: (i) ρ (networks' endogeneity) increases (see sub-figures (c)-(f) in Figure 3), or (ii) the sample size increases (see Figure 4). Finally, in design **(II)** (Scale-free distribution), while the asymptotic tests perform similarly in the two partitions (and also outperform their bootstrap counterparts in most cases), the power of the bootstrap tests is lower with partition $\mathbf{P}_n^{(4)}$ than with $\mathbf{P}_n^{(3)}$. The power gain from using $\mathbf{P}_n^{(3)}$ over $\mathbf{P}_n^{(4)}$ can even be substantial, especially with the bootstrap test $\mathbf{S}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ (see sub-figures (d) and (f) in Figure 5). Again, partition $\mathbf{P}_n^{(3)}$ is closer to Mann and Wald's (1942) and Williams's (1950) recommendation than partition $\mathbf{P}_n^{(4)}$.

Clearly, although bootstrapping has an overall good performance in terms of size control irrespective of which partition in $\mathscr{P}_A$ is used, our Monte Carlo results suggest that using the partition that is closer to equalizing the expected number in cells can results in a substantial power gain. Therefore, our recommendation is to follow this rule upon adjusting for the form of the distribution of the degrees, as discussed in (2) - (6) of Section 3.2.

As we are bootstrapping a quantity which is asymptotically pivotal, one might expect the bootstrap to provide the usual asymptotic refinement. However, the size and power properties illustrated here do not seem to provide a clear benefit in using the bootstrap over the asymptotic test, suggesting that any such refinement may be quite minor. One possible reason for this is that the examples considered here consider distributions in which the number of cells in the partition is quite small, allowing the vector v to approach multivariate normality relatively quickly.



(a) $\mathbf{W}_m$ with $\rho = 0$.

(b) $\mathbf{S}_m$ with $\rho = 0$.

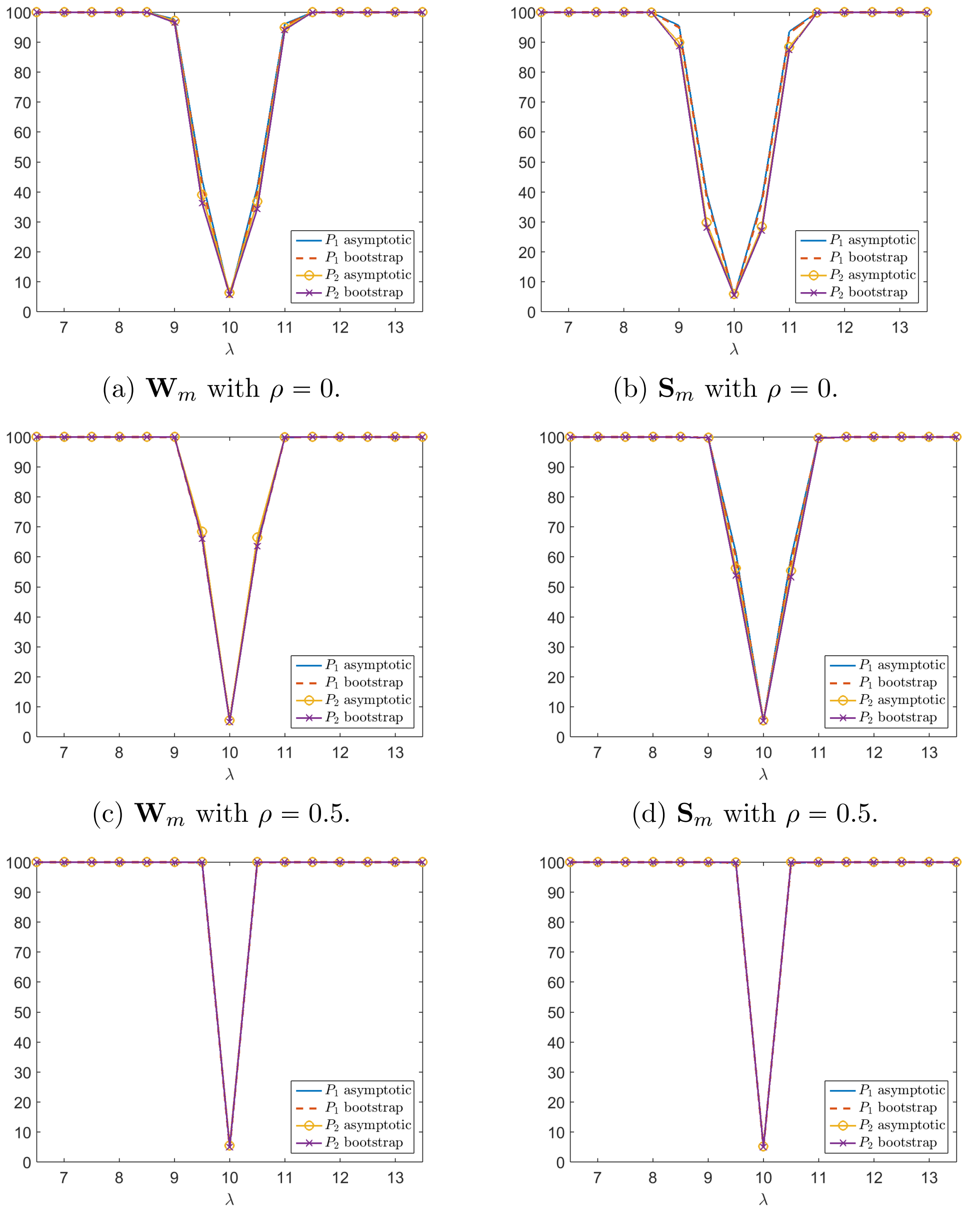
(c) $\mathbf{W}_m$ with $\rho = 0.5$.

(d) $\mathbf{S}_m$ with $\rho = 0.5$.

(e) $\mathbf{W}_m$ with $\rho = 0.9$.

(f) $\mathbf{S}_m$ with $\rho = 0.9$.

Figure 3: Power curves for $n = 100$



(a) $\mathbf{W}_m$ with $\rho = 0$.

(b) $\mathbf{S}_m$ with $\rho = 0$.

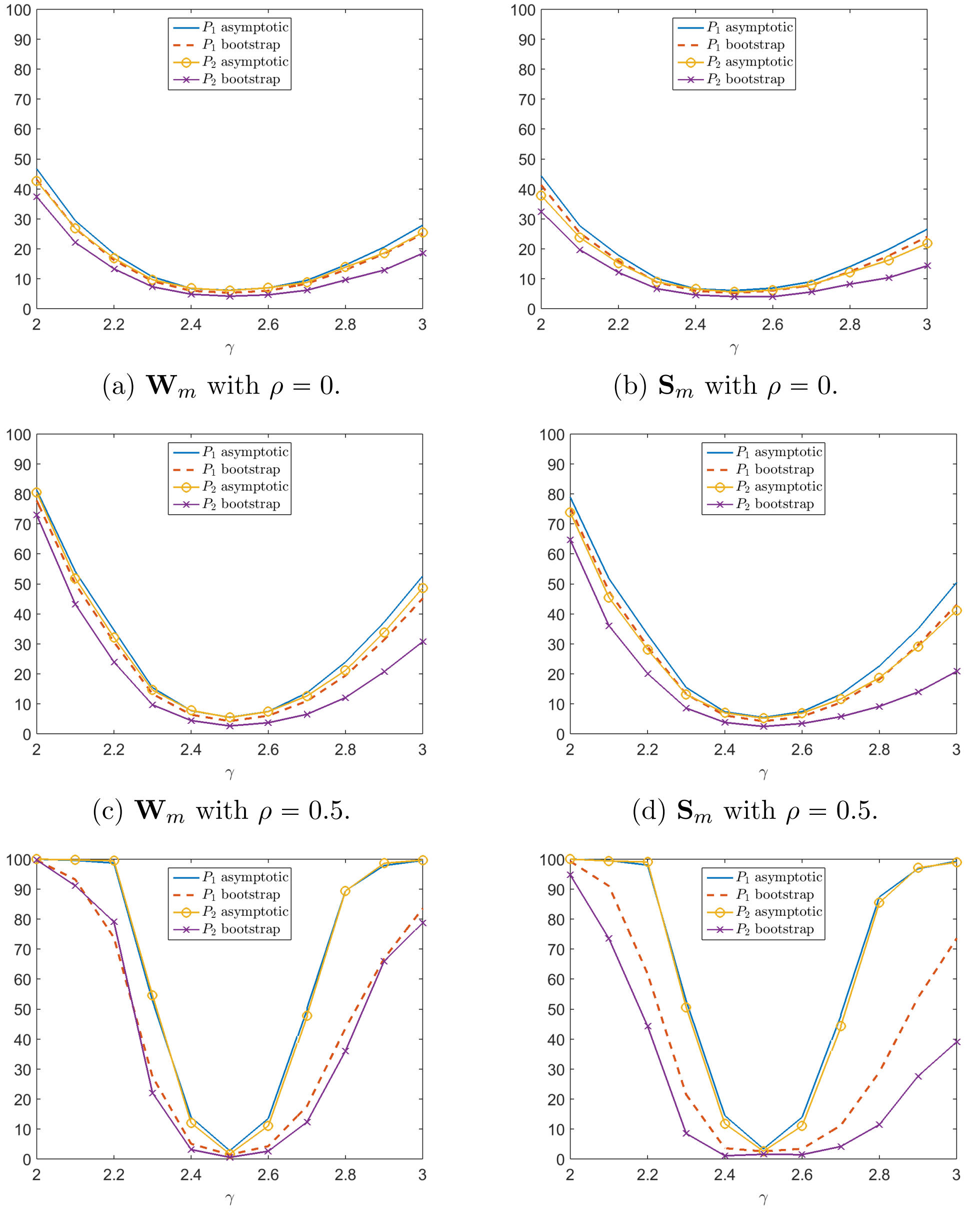
(c) $\mathbf{W}_m$ with $\rho = 0.5$.

(d) $\mathbf{S}_m$ with $\rho = 0.5$.

(e) $\mathbf{W}_m$ with $\rho = 0.9$.

(f) $\mathbf{S}_m$ with $\rho = 0.9$.

Figure 4: Power curves for $n = 500$



(a) $\mathbf{W}_m$ with $\rho = 0$.

(b) $\mathbf{S}_m$ with $\rho = 0$.

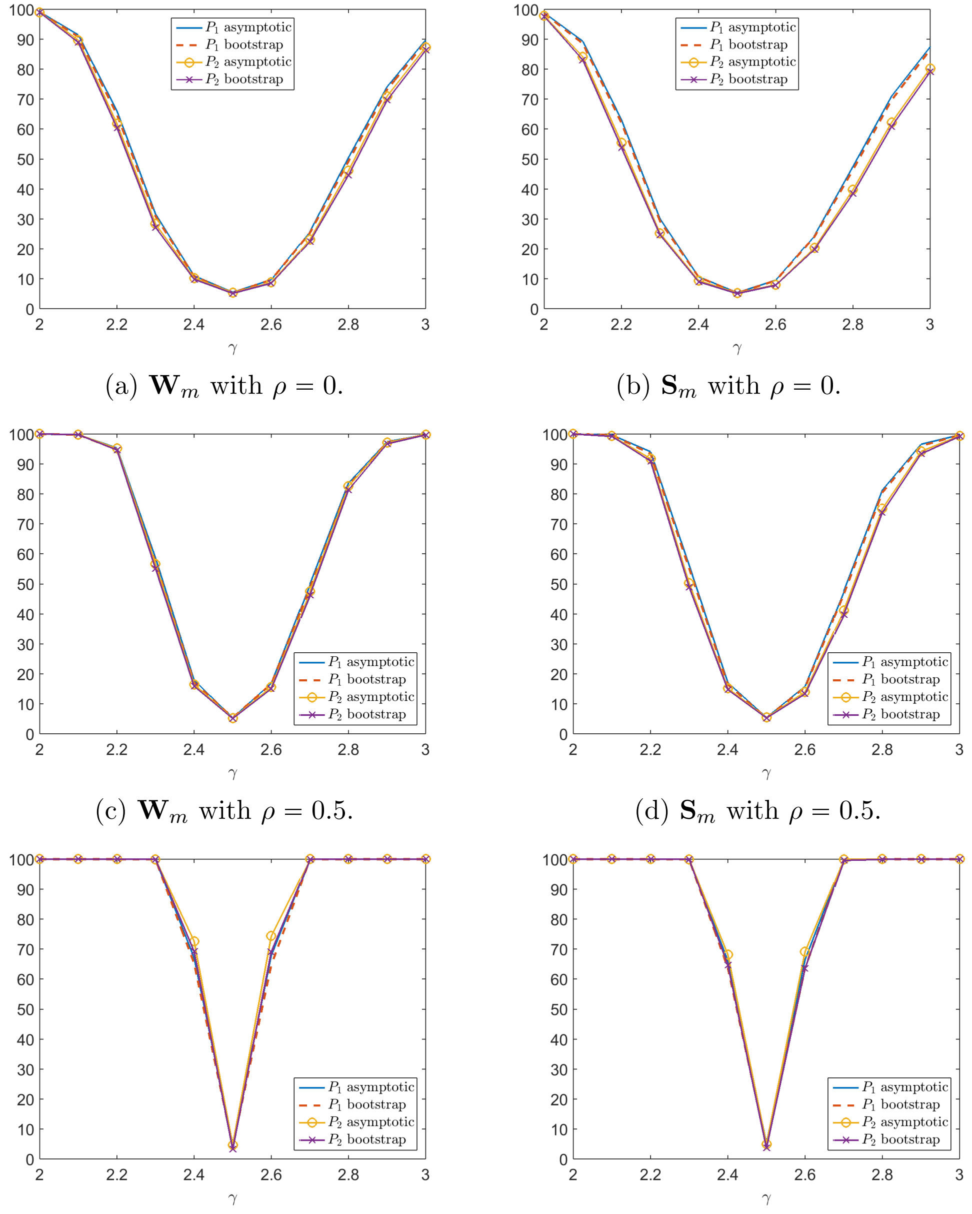
(c) $\mathbf{W}_m$ with $\rho = 0.5$.

(d) $\mathbf{S}_m$ with $\rho = 0.5$.

(e) $\mathbf{W}_m$ with $\rho = 0.9$.

(f) $\mathbf{S}_m$ with $\rho = 0.9$.

Figure 5: Power curves for $n = 100$



(a) $\mathbf{W}_m$ with $\rho = 0$.

(b) $\mathbf{S}_m$ with $\rho = 0$.

(c) $\mathbf{W}_m$ with $\rho = 0.5$.

(d) $\mathbf{S}_m$ with $\rho = 0.5$.

(e) $\mathbf{W}_m$ with $\rho = 0.9$.

(f) $\mathbf{S}_m$ with $\rho = 0.9$.

Figure 6: Power curves for $n = 500$

6. Empirical Illustration

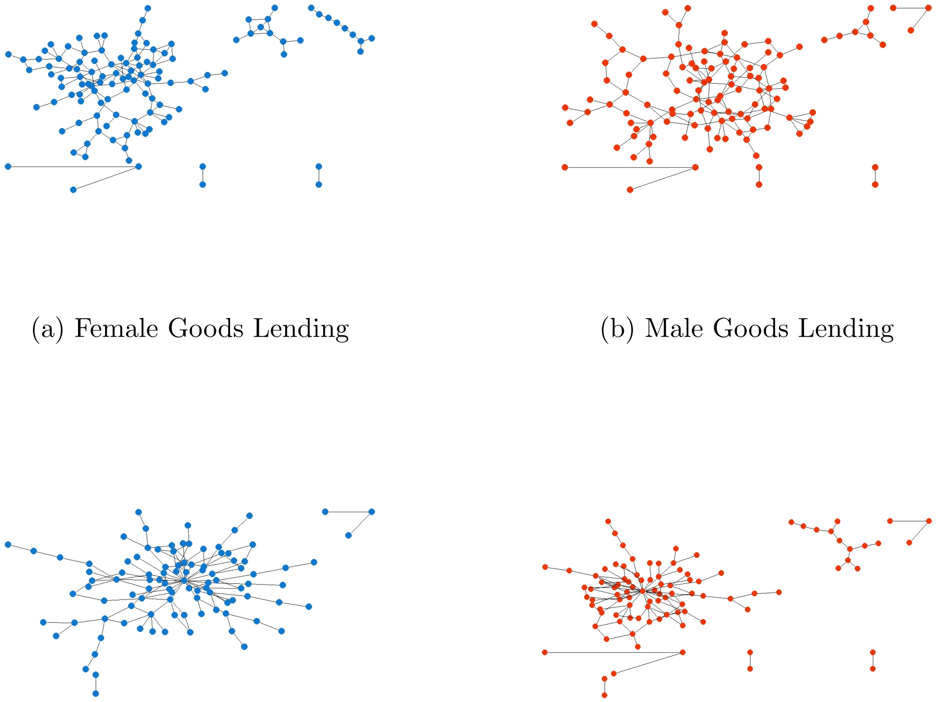
Rosenzweig and Stark (1989) illustrate the strategic role that women play in smoothing consumption between villages whose income shocks are negatively correlated. In this application, we investigate whether such a role exists for sharing risk between households in rural India. In particular, we focus on testing gender differences across risk sharing networks by using the stochastic dominance criteria. Bramoulle and Kranton (2007) characterize the conditions that insure the existence of an aggregate strictly increasing (and even concave) social welfare function in risk sharing networks, meaning that these networks could be ranked in terms of *social efficiency* by applying the stochastic dominance criteria in Definition 1.

We use the data set from Banerjee et al. (2012, 2013) and Jackson et al. (2012) that comprise a random sample of households from 75 different villages in southern India. We pool the sub-samples from these villages to obtain one sample. The underlying assumption here is that the 75 subsamples are independent across villages, but not at the household level. Each village contains on average 223 households with approximately half being sampled. Each member of a surveyed household was asked to identify members of the village with whom they engaged in a particular relationship, such as whose home they visit or with whom they go to temple. Additionally, a census on the socioeconomic characteristics– such as age, gender, religion, etc– of households was used to complete the data set; see Banerjee et al. (2012, 2013) and Jackson et al. (2012) for a detailed description of the data.

To identify risk sharing behavior we use data on the following questions:

Who would come to you if he (or she) needed to borrow kerosene or rice? Who do you trust enough that if he (or she) needed to borrow 50 rupees for a day you would lend it to him (or her)? We construct female and male networks for each of the *goods lending* and *money lending* relationships as follows. We remove from the sample any person who does not name at least one connection, as it is difficult to distinguish non-response from having zero connections. We also remove any person under the age of 18. Of the remaining observations, we omit any household which does not contain at least one man and one woman. The networks are then constructed with a node representing each household. In the *female money lending network*, there is a directed link from household i to household i' if any woman in household i has reported that she would lend money to any member (male or female) of household i', and similarly for the *male money lending network*. This means that the male and female networks have the same set of households as *nodes* and the gender corresponding to the network determines the set of *directed links*. The goods lending networks are constructed similarly. As an illustration, Figure 7 shows these networks within the households of village 1 in the data.

Figure 7: Risk Sharing Networks for Village 1



(a) Female Goods Lending

(b) Male Goods Lending

(c) Female Money lending

(d) Male Money Lending

As outlined above, we conduct the tests using the pooled sample of all 75 villages. The pooled sample has size $n = 5924$ households in *goods lending networks*, and $n = 5656$ households in *money lending networks*. Table 4 summarizes the out-degree distributions of these networks as well as the correlations between *male and female networks* for both *goods lending* and *money lending*. As seen, the correlation between *male and female net-*

works is not small: 0.55 (for goods lending) and 0.46 (for money lending). Furthermore, in each case (goods lending and money lending) the degree distributions of both male and female networks are closer to the degree distribution of a Poisson random graph than that of a scale-free network (see Figure 2). From Sections 3.2-5, we use the following admissible partition with $k = 5$ based on Table 4:

$$\mathbf{P}_n^{(5)}(\mathbf{I}) = \{\mathbf{I}_l\}_{l=1}^{5}, \ \mathbf{I}_l = \{l\} \text{ for } l = 1, \ldots, 4 \text{ and } \mathbf{I}_5 = \{5+\}. \tag{19}$$

Table 4: Empirical Degree Distributions

	Goods		Money	
Degree	Male	Female	Male	Female
1	527	426	962	1012
2	2554	2133	2653	2509
3	1801	1831	1270	1263
4	734	1014	460	564
5	172	306	164	194
6	94	136	82	69
7	32	46	39	21
8	7	19	17	14
9	0	5	3	3
10	2	6	2	5
11	0	2	1	1
12	0	0	0	1
13	0	0	2	0
14	1	0	1	0
Obs.	5924	5924	5656	5656
Correlation	0.55		0.46	

In both the goods lending and money lending networks, we test whether the *female network* first- and second-order stochastically dominates the *male network*. The tests are run at the 1% and 5% nominal levels, and the

Table 5: Stochastic dominance between female and male networks

Goods lending networks

	$\alpha = 0.01$		$\alpha = 0.05$	
Statistics ↓ m →	1	2	1	2
$\mathbf{W}_m$	310.08	310.08	310.08	310.08
$\chi^2_4(\alpha)$	13.28	13.28	9.49	9.49
$c^*_{\mathbf{W}_m}(\alpha)$	11.49	12.36	9.32	10.09
$\mathbf{S}_m$	16.92	17.01	16.92	17.01
$z_4(\alpha)$	3.02	3.02	2.49	2.49
$c^*_{\mathbf{S}_m}(\alpha)$	3.17	2.93	2.64	2.60

Money lending networks

	$\alpha = 0.01$		$\alpha = 0.05$	
Statistics ↓ m →	1	2	1	2
$\mathbf{W}_m$	19.29	19.29	19.29	19.29
$\chi^2_4(\alpha)$	13.28	13.28	9.49	9.49
$c^*_{\mathbf{W}_m}(\alpha)$	15.60	16.73	11.07	8.80
$\mathbf{S}_m$	2.92	2.59	2.92	2.59
$z_4(\alpha)$	3.02	3.02	2.49	2.49
$c^*_{\mathbf{S}_m}(\alpha)$	3.05	3.38	2.59	2.47

† $\chi^2_4(\alpha)$ and $z_4(\alpha)$ are the $(1-\alpha)^{th}$ quantiles of a chi-squared distributed random variable with 4 degrees of freedom a $SMM(4,\infty)$-distributed random variable respectively.

‡ $c^*_{\mathbf{W}_m}(\alpha)$ and $c^*_{\mathbf{S}_m}(\alpha)$ are the $(1-\alpha)^{th}$ critical values of the bootstrap statistics $\mathbf{W}^*_m$ and $\mathbf{S}^*_m$ respectively. Note that $c^*_{\mathbf{W}_1}$ theoretically equals $c^*_{\mathbf{W}_2}$ as $\mathbf{W}_m$ is invariant to the choice of m. The differ here as they are constructed from different bootstrap samples.

bootstrap statistics critical values are approximated using $B = 199$ pseudo-samples. The results are displayed in table 5. For goods lending, both the asymptotic and bootstrap tests are in favor of the first- and second-order stochastic dominance of the *female network* at the 1% and 5% nominal levels. However, for money lending, we could only find evidence of the first- and second-order dominance of the female network at the 5% nominal level. At the 1% nominal level, neither network dominates the other using both the asymptotic and bootstrap tests. These results suggest that women overall tend to form denser risk sharing networks than do men, especially for goods lending. One possible explanation for this might be a higher average risk aversion among women, as documented by Borghans et al. (2009).

7. Conclusion

This paper has illustrated how stochastic dominance criteria can be used to rank networks in terms of social efficiency, and developed statistical tests for assessing these criteria. The tests proposed can be seen as a generalization of the Pearson-type and the studentized maximum modulus (SMM)-type statistics usually employed for assessing stochastic dominance criteria in the literature on income distributions, poverty and inequality measures. Our statistics differ from the prior literature not only through a correction to account for the correlation between the degree distributions of networks, but also their direct dependence on partitioning into classes. We show that a combination of the modified Pearson- and SMM-type statistics into a single decision rule is necessary to inform us on whether *stochastic dominance* holds or not, once equality between the degree distributions of the networks is rejected. As these statistics often depend on the way class intervals are allocated, controlling for type-I error uniformly over the set of all *admissible class allocations*[20] is important for the asymptotic results to give a good approximation of their empirical size to the nominal level.

We provide an analysis of both the size and power properties of the tests. On level control, we establish uniform convergence of their empirical size to the nominal level when the usual asymptotic chi-square and SMM critical values are applied. On power, we show that test consistency holds no matter which *admissible* partition is used. Finally, we provide a bootstrap method that enhances the finite-sample performance of the tests. We establish uniform consistency of the bootstrap for both the proposed

[20] By *admissible class allocation* or *admissible partition*, we mean a partition in which the minimum expected number in each cell is at least 5.

Pearson- and SMM-tests irrespective of whether the null hypothesis holds or not. We present a Monte Carlo experiment that confirms our theoretical findings. Using the data set of Jackson et al. (2012) and Banerjee et al. (2012, 2013), the proposed tests were illustrated through an investigation of households' *risk sharing networks* across 75 villages in rural India. Both the goods lending and money lending networks were considered, and the gender difference within each network was our main focus. Our results suggested that women within these villages overall tend to form denser *risk sharing networks* than do men, especially for goods lending.

Appendix A. Appendix: Proofs

In order to establish the proofs of the lemmata and theorems of the main text, it is useful to state some basic convergence of covariance matrices $\widehat{\Sigma}_j$, $j \in \{g, g'\}$, $\widehat{\Sigma}_{gg'}$, and $\widehat{\Omega}_m$ given in (9) - (12).

Lemma 4. *Suppose that Assumption 1 holds. For any $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$, we have:*

$$
\begin{aligned}
(i.)\ \underset{n\to\infty}{p\lim}(\widehat{\Sigma}_j) &= \Sigma_j := \begin{pmatrix} p_{j\cdot 1}(1-p_{j\cdot 1}) & -p_{j\cdot 1}p_{j\cdot 2} & \dots & -p_{j\cdot 1}p_{j\cdot k} \\ -p_{j\cdot 2}p_{j\cdot 1} & p_{j\cdot 2}(1-p_{j\cdot 2}) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ -p_{j\cdot k}p_{j\cdot 1} & -p_{j\cdot k}p_{j\cdot 2} & \dots & p_{j\cdot k}(1-p_{j\cdot k}) \end{pmatrix} \forall\, j \in \{g, g'\}, \\
(ii.)\ \underset{n\to\infty}{p\lim}(\widehat{\Sigma}_{gg'}) &= \Sigma_{gg'} := \begin{pmatrix} p_{gg'\cdot 11} - p_{g\cdot 1}p_{g'\cdot 1} & p_{gg'\cdot 12} - p_{g\cdot 1}p_{g'\cdot 2} & \dots & p_{gg'\cdot 1k} - p_{g\cdot 1}p_{g'\cdot k} \\ p_{gg'\cdot 21} - p_{g\cdot 2}p_{g'\cdot 1} & p_{gg'\cdot 22} - p_{g\cdot 2}p_{g'\cdot 2} & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ p_{gg'\cdot k1} - p_{g\cdot k}p_{g'\cdot 1} & p_{gg'\cdot k2} - p_{g\cdot k}p_{g'\cdot 2} & \dots & p_{gg'\cdot kk} - p_{g\cdot k}p_{g'\cdot k} \end{pmatrix}, \\
(iii.)\ \underset{n\to\infty}{p\lim}(\widehat{\Omega}_m) &= \Omega_m := \mathbf{T}^m(\Sigma_g + \Sigma_{g'} - \Sigma_{gg'} - \Sigma'_{gg'})\mathbf{T}'^m.
\end{aligned}
$$

Proof of Lemma 4. $(i.)$ Suppose that Assumption 1 holds and let $\mathbf{P}_n^{(k)}(\mathbf{I}) = \{\mathbf{I}_l\}_{l=1}^k \in \mathscr{P}_A$. From the i.i.d. sampling, it follows that $\hat{p}_{j\cdot l} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(d_{j\cdot i} \in \mathbf{I}_l) \xrightarrow{p} \mathbb{E}(d_{j\cdot i}) = p_{j\cdot il} \equiv p_{j\cdot l}$ for all $(j, l) \in \{g, g'\} \times \{1, \dots, k\}$. It is clear from (9) that $\widehat{\Sigma}_j \xrightarrow{p} \Sigma_j$ for all $j \in \{g, g'\}$. The proof of $(ii.)$ follows the same steps and $(iii.)$ is implied by $(i.)$ and $(ii.)$.

Proof of Lemma 1. Let $\mathbf{P}_n^{(k)}(\mathbf{I}) = \{\mathbf{I}_l\}_{l=1}^k \in \mathscr{P}_A$ and define

$$
\hat{p} \;=\; [\hat{p}'_g \;:\; \hat{p}'_{g'}]',\ p = [p'_g \;:\; p'_{g'}]', \tag{A.1}
$$

where $\hat{p}_g = [\hat{p}_{g\cdot 1}, \dots, \hat{p}_{g\cdot k}]' : k \times 1$, $\hat{p}_{g'} = [\hat{p}_{g'\cdot 1}, \dots, \hat{p}_{g'\cdot k}]' : k \times 1$, $p_g = [p_{A\cdot 1}, \dots, p_{A\cdot k}]' : k \times 1$, and $p_{g'} = [p_{B\cdot 1}, \dots, p_{B\cdot k}]' : k \times 1$, so both $\hat{p}$ and p are

$2k \times 1$ vectors obtained by stacking $\hat{p}_g$ and $\hat{p}_{g'}$ together (for $\hat{p}$) and p_g and $p_{g'}$ together (for p). From (7) - (8), we have $\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} u_{j\cdot i}$ and for each $j \in \{g, g'\}$, $u_{j\cdot i}$, $i = 1, \ldots, n$ are i.i.d. multinomial random variables with parameter $p_j = \mathbb{E}(u_{j\cdot i})$ under Assumption 1. Therefore, by the multivariate central limit theorem (MVCLT), we have:

$$\sqrt{n}(\hat{p} - p) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} u_{g\cdot i} - \mathbb{E}(u_{g\cdot i}) \\ u_{g'\cdot i} - \mathbb{E}(u_{g'\cdot i}) \end{bmatrix} \xrightarrow{d} N\left(0,\ \Sigma_p\right), \tag{A.2}$$

where $\Sigma_p = Avar\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} u_{g\cdot i} - \mathbb{E}(u_{g\cdot i}) \\ u_{g'\cdot i} - \mathbb{E}(u_{g'\cdot i}) \end{bmatrix}\right) = \begin{bmatrix} \Sigma_g & \Sigma_{gg'} \\ \Sigma'_{gg'} & \Sigma_{g'} \end{bmatrix}$, Σ_j and $\Sigma_{gg'}$ are the limits in Lemma 4. Now, let I_k be the identity matrix of order k. By noting that

$$\begin{aligned} \begin{bmatrix} I_k & -I_k \end{bmatrix}\sqrt{n}(\hat{p} - p) &= \begin{bmatrix} I_k & -I_k \end{bmatrix}\sqrt{n}\begin{bmatrix} \hat{p}_g - p_g \\ \hat{p}_{g'} - p_{g'} \end{bmatrix} \\ &= \sqrt{n}[(\hat{p}_g - \hat{p}_{g'}) - (p_g - p_{g'})], \end{aligned} \tag{A.3}$$

it is straightforward to see that $\sqrt{n}[(\hat{p}_g - \hat{p}_{g'}) - (p_g - p_{g'})] \xrightarrow{d} N\left[0,\ \Sigma_g + \Sigma_{g'} - (\Sigma_{gg'} + \Sigma'_{gg'})\right]$ from (A.2). This completes the proof of Lemma 1.

Proof of Lemma 2. Suppose that Assumption 1 holds and let $\mathbf{P}_n^{(k)}(\mathbf{I}) = \{\mathbf{I}_l\}_{l=1}^{k} \in \mathscr{P}_A$.

(a) Assume first that H_{0m} holds, i.e., $p_g = p_{g'}$. We focus on the statistic $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$. The proof for $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ can easily be adapted from Stoline and Ury (1979). From Lemmas 1 and 4, along with the expression of $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))$ in (12), it is straightforward to see that $\sqrt{n}\mathbf{T}^m[(\hat{p}_g - \hat{p}_{g'}) - (p_g - p_{g'})] \overset{H_{0m}}{=} \sqrt{n}\mathbf{T}^m(\hat{p}_g - \hat{p}_{g'}) \xrightarrow{d} \psi_m \sim N\left(0,\ \Omega_m\right)$ so that we get

$$\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{d} \psi'_m \Omega_m^{-} \psi_m, \tag{A.4}$$

where $\Omega_m = \Sigma_g + \Sigma_{g'} - (\Sigma_{gg'} + \Sigma'_{gg'})$, and Ω_m^- is the generalized inverse of Ω_m. As $rank(\Omega_m) = k - 1$, there exists [see Dhrymes (1978, Proposition 3.5)] a diagonal matrix D_{k-1} whose diagonal elements are the nonzero eigenvalues of Ω_m (in decreasing order of magnitude), and a $k \times (k-1)$ matrix P_{k-1} whose columns are the (orthogonal) eigenvectors corresponding to the nonzero roots of Ω_m, such that

$$\Omega_m = P_{k-1} D_{k-1} P'_{k-1} \text{ and } \Omega_m^- = P_{k-1} D_{k-1}^{-1} P'_{k-1}. \tag{A.5}$$

Hence, we have: $\psi'_m \Omega_m^- \psi_m = \psi'_m P_{k-1} D_{k-1}^{-1} P'_{k-1} \psi_m = \bar{\psi}'_m D_{k-1}^{-1} \bar{\psi}_m$ from the last identity in (A.5), where $\bar{\psi}_m = P'_{k-1}\psi_m$. Since $\psi_m \sim N\left(0,\ \Omega_m\right)$, we have $\bar{\psi}_m \sim N\big(0,\ P'_{k-1}\Omega_m P_{k-1}\big) = N\big(0,\ P'_{k-1}P_{k-1}D_{k-1}P'_{k-1}P_{k-1}\big) = N\left(0,\ D_{k-1}\right)$ from the first identity in (A.5), where $P'_{k-1}P_{k-1} = I_{k-1}$. Therefore, $D_{k-1}^{-1/2}\bar{\psi}_m \sim N\left(0,\ I_{k-1}\right)$ so that $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{d} \bar{\psi}'_m D_{k-1}^{-1} \bar{\psi}_m \sim \chi^2(k - 1)$, as stated.

(b) Assume now that H_{1m} or H_{2m} is true. Hence, we have $p_g - p_{g'} \neq 0$ so that $\hat{v}_m \xrightarrow{p} v_m = \mathbf{T}^m(p_g - p_{g'}) \neq 0$. Furthermore, as $\widehat{\Omega}_m \xrightarrow{p} \Omega_m$, it is clear that $\hat{v}'_m \widehat{\Omega}_m^- \hat{v}_m \xrightarrow{p} (p_g - p_g)'\mathbf{T}^{m\prime}\Omega_m^-\mathbf{T}^m(p_g - p_{g'}) > 0$ because $rank(\Omega_m^-) = k - 1$. Therefore, we find $\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) = n\hat{v}'_m \widehat{\Omega}_m^- \hat{v}_m \xrightarrow{p} +\infty$. Similarly, we can see that $\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \xrightarrow{p} +\infty$. This completes the proof of Lemma 2.

Proof of Theorem 1. (a) Suppose first that H_{0m} holds. Since $\mathscr{P}_A$ is a discrete and finite set of collection of partitions $\mathbf{P}_n^{(k)}(\mathbf{I})$, the sequence of probabilities $\alpha_{1,n}^{(k)}[\mathbf{P}_n^{(k)}(\mathbf{I}), \mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))] = \mathbb{P}[\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > \chi^2_{k-1}(\alpha)] \in [0, 1]$ and $\alpha_{2,n}^{(k)}[\mathbf{P}_n^{(k)}(\mathbf{I}), \mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))] = \mathbb{P}[\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) > z_{k-1}(\alpha)] \in [0, 1]$ can be *ordered* for all possible collections $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_A$. Therefore, there are sequences $\widetilde{\mathbf{P}}_n^{(k)}, \breve{\mathbf{P}}_n^{(k)} \in \mathscr{P}_A$ and subsequences $\{\pi_n : n \geqslant 1\}, \{\breve{\pi}_n : n \geqslant 1\}$ of

$\{n: \ n \geqslant 1\}$ such that

$$\begin{aligned}\limsup_{n\to\infty}\sup_{\mathscr{P}_A}\alpha_{1,n}^{(k)}[\mathbf{P}_n^{(k)}(\mathbf{I}),\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))] &:= \limsup_{n\to\infty}\sup_{\mathscr{P}_A}\mathbb{P}[\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))>\chi^2_{k-1}(\alpha)]\\ &= \limsup_{n\to\infty}\mathbb{P}[\mathbf{W}_m(\widetilde{\mathbf{P}}_n^{(k)})>\chi^2_{k-1}(\alpha)]\\ &= \lim_{n\to\infty}\mathbb{P}[\mathbf{W}_m(\widetilde{\mathbf{P}}_{\pi_n}^{(k)})>\chi^2_{k-1}(\alpha)]. \quad (A.6)\\ \limsup_{n\to\infty}\sup_{\mathscr{P}_A}\alpha_{2,n}^{(k)}[\mathbf{P}_n^{(k)}(\mathbf{I}),\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))] &:= \limsup_{n\to\infty}\sup_{\mathscr{P}_A}\mathbb{P}[\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))>z_{k-1}(\alpha)]\\ &= \limsup_{n\to\infty}\mathbb{P}[\mathbf{S}_m(\breve{\mathbf{P}}_n^{(k)})>z_{k-1}(\alpha)]\\ &= \lim_{n\to\infty}\mathbb{P}[\mathbf{S}_m(\breve{\mathbf{P}}_{\breve{\pi}_n}^{(k)})>z_{k-1}(\alpha)]. \quad (A.7)\end{aligned}$$

But from Lemma 2-(a), we have $\lim_{n\to\infty}\mathbb{P}[\mathbf{W}_m(\widetilde{\mathbf{P}}_{\pi_n}^{(k)})>\chi^2_{k-1}(\alpha)]=\mathbb{P}[\chi^2_{k-1}>\chi^2_{k-1}(\alpha)]=\alpha$ and $\lim_{n\to\infty}\mathbb{P}[\mathbf{S}_m(\breve{\mathbf{P}}_{\breve{\pi}_n}^{(k)})>z_{k-1}(\alpha)]=\mathbb{P}[SMM(k,\infty)>z_{k-1}(\alpha)]=\alpha$. Using (A.6)-(A.7), we get:

$$\limsup_{n\to\infty}\sup_{\mathscr{P}_A}\mathbb{P}[\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))>\chi^2_{k-1}(\alpha)]=\alpha \text{ and } \limsup_{n\to\infty}\sup_{\mathscr{P}_A}\mathbb{P}[\mathbf{S}_m(\mathbf{P}_n^{(k)}(\mathbf{I}))>z_{k-1}(\alpha)]=\alpha.$$

(b) Under H_{1m} or H_{2m}, the results follow immediately from Lemma 2-(b).

Proof of Lemma 3. We prove the results for $\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$. The proof for $\mathbf{S}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ can be constructed in a similar way. First, we can write the bootstrap statistic $\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ as

$$\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \;=\; n\tilde{v}_m^{*\prime}\widehat{\Omega}_m^{*-}\tilde{v}_m^*=n(\hat{v}_m^*-\hat{v}_m)'\widehat{\Omega}_m^{*-}(\hat{v}_m^*-\hat{v}_m). \quad (A.8)$$

(a) Suppose first that H_{0m} holds and let $S_m^*=\sqrt{n}(\hat{v}_m^*-\hat{v}_m)$. We can express S_m^* as:

$$S_m^*=\sum_{i=1}^n R_{m,i}^*, \text{ where } R_{m,i}^*=\frac{1}{\sqrt{n}}\mathbf{T}^m\Big[(d_{g\cdot i}^*-d_{g'\cdot i}^*)-\frac{1}{n}\sum_{i=1}^n(u_{g\cdot i}-u_{g'\cdot i})\Big].$$

Moreover, from the i.i.d. sampling under $\mathbb{P}^*$, we have $\mathbb{E}^*(d_{g\cdot i}^*-d_{g'\cdot i}^*)=\frac{1}{n}\sum_{i=1}^n(u_{g\cdot i}-u_{g'\cdot i})$, so that $R_{m,i}^*$ can be expressed as $R_{m,i}^*=\frac{1}{\sqrt{n}}\mathbf{T}^m\big[d_{g\cdot i}^*-d_{g'\cdot i}^*-\mathbb{E}^*(d_{g\cdot i}^*-d_{g'\cdot i}^*)\big]$, i.e., $\{R_{m,i}^*\}_{i=1}^n$ are also i.i.d under $\mathbb{P}^*$. We want to verify the conditions of the Liapunov Central Limit Theorem for S_m^*.

(a) By definition, it is straightforward to see that $\mathbb{E}^*(R_{m,i}^*)=0$.

(b) $\mathbb{E}^*(R^{*^2}_{m,i}) = var^*(R^*_{m,i}) = n^{-1}\widehat{\Omega}_m < \infty\ a.s.$

(c) Finally, we need to show that $\lim\limits_{n\to\infty}\sum_{i=1}^n \mathbb{E}^*[\|R^*_{m,i}\|^{2+\delta}] = 0\ a.s.$ for some $\delta > 0$. We have:

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}^*[\|R^*_{m,i}\|^{2+\delta}] &\leqslant cn^{-\frac{\delta}{2}}n^{-1}\sum_{i=1}^n \mathbb{E}^*\Big[\|\mathbf{T}^m(d^*_{g\cdot i}-d^*_{g'\cdot i})\|^{2+\delta} + \|\frac{1}{n}\sum_{i=1}^n \mathbf{T}^m(u_{g\cdot i}-u_{g'\cdot i})\|^{2+\delta}\Big] \\ &= cn^{-\frac{\delta}{2}}\mathbb{E}^*\Big[\|\mathbf{T}^m(d^*_{g\cdot i}-d^*_{g'\cdot i})\|^{2+\delta}\Big] + cn^{-\frac{\delta}{2}}\Big\|\frac{1}{n}\sum_{i=1}^n \mathbf{T}^m(u_{g\cdot i}-u_{g'\cdot i})\Big\|^{2+\delta}\end{aligned}$$

for a large enough constant $c \in \mathbb{R}^+$.

First, we have $\frac{1}{n}\sum_{i=1}^n \mathbf{T}^m(u_{g\cdot i}-u_{g'\cdot i}) \xrightarrow{p} \mathbf{T}^m(p_g - p_{g'}) = v_m = 0$ under Assumption 1 and H_{0m}. So, the second term of the last equality in the above equation is such that $cn^{-\frac{\delta}{2}}\Big\|\frac{1}{n}\sum_{i=1}^n \mathbf{T}^m(u_{g\cdot i}-u_{g'\cdot i})\Big\|^{2+\delta} \xrightarrow{p} 0$ since $cn^{-\frac{\delta}{2}} \to 0$ when $n \to \infty$. For the first term, we note that $\mathbb{E}^*\big[\|\mathbf{T}^m(d^*_{g\cdot i}-d^*_{g'\cdot i})\|^{2+\delta}\big] \xrightarrow{p^*} \Big\|\frac{1}{n}\sum_{i=1}^n \mathbf{T}^m(u_{g\cdot i}-u_{g'\cdot i})\Big\|^{2+\delta}$ and we know that $\Big\|\frac{1}{n}\sum_{i=1}^n \mathbf{T}^m(u_{g\cdot i}-u_{g'\cdot i})\Big\|^{2+\delta} \xrightarrow{p} \big\|\mathbf{T}^m(p_g - p_{g'})\big\|^{2+\delta} = \|v_m\|^{2+\delta} = 0$ when H_{0m} holds. So, we get $cn^{-\frac{\delta}{2}}\mathbb{E}^*\big[\|\mathbf{T}^m(d^*_{g\cdot i}-d^*_{g'\cdot i})\|^{2+\delta}\big] \xrightarrow{p} 0\ a.s.$ As a result, we have $\lim\limits_{n\to\infty}\sum_{i=1}^n \mathbb{E}^*[\|R^*_{m,i}\|^{2+\delta}] = 0\ a.s.$ as required.

Since $\widehat{\Omega}^*_m - \widehat{\Omega}_m \mid \mathscr{D}_n \xrightarrow{a.s.} 0$, $\widehat{\Omega}_m \xrightarrow{p} \Omega_m$, and the conditions of the Liapunov CLT are satisfied,we have

$$S^*_m \mid \mathscr{D}_n \ \xrightarrow{d}\ \psi_m \sim N(0,\, \Omega_m)\ a.s.$$

Now, we want to show that $\mathbf{W}^*_m(\mathbf{P}^{(k)}_n(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{d} \chi^2(k-1)\ a.s.$ for any $\mathbf{P}^{(k)}_n(\mathbf{I}) \in \mathscr{P}_{_A}$. From (A.8) and the fact that $\widehat{\Omega}^*_m \mid \mathscr{D}_n \xrightarrow{p} \Omega_m\ a.s.$, it is straightforward to see that

$$\mathbf{W}^*_m(\mathbf{P}^{(k)}_n(\mathbf{I})) \mid \mathscr{D}_n \ = \ S^{*'}_m\widehat{\Omega}^{*^-}_m S^*_m \mid \mathscr{D}_n \xrightarrow{d} \psi'_m\Omega^-_m\psi_m\ a.s. \tag{A.9}$$

Since we have $\psi'_m\Omega^-_m\psi_m \sim \chi^2(k-1)$ by Lemma 2, it is clear that $\mathbf{W}^*_m(\mathbf{P}^{(k)}_n(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{d} \chi^2(k-1)\ a.s.$ for all $\mathbf{P}^{(k)}_n(\mathbf{I}) \in \mathscr{P}_{_A}$, as stated.

(b) Suppose now that H_{0m} fails, i.e., H_{1m} or H_{2m} holds. It is easy to see from the proof in (a) that $\frac{1}{\sqrt{n}}S^{*'}_m \mid \mathscr{D}_n \xrightarrow{p} v_m\ a.s.$, $\widehat{\Omega}^*_m \mid \mathscr{D}_n \xrightarrow{p} \Omega_m\ a.s.$

so that $\frac{1}{n}\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{a.s.} v_m'\Omega_m^- v_m > 0$ because $v_m \neq 0$ under H_{1m} or H_{2m}. Therefore, we have $\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \mid \mathscr{D}_n \xrightarrow{p} +\infty$ $a.s.$ under H_{1m} or H_{2m} for any $\mathbf{P}_n^{(k)}(\mathbf{I}) \in \mathscr{P}_{_A}$, as required.

Proof of Theorem 2. As in Lemma 3, we will prove the results for $\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$. The proof for $\mathbf{S}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ can be constructed in a similar way.

(a) Suppose first that H_{0m} holds. We know from Lemma 3 that $\widehat{\Omega}_m^* - \widehat{\Omega}_m \mid \mathscr{D}_n \xrightarrow{a.s.} 0$ and $\widehat{\Omega}_m$ has rank $k-1$ by construction. Hence, $\widehat{\Omega}_m^*$ also has rank $k-1$ $a.s.$ Therefore, from Dhrymes (1978, Proposition 3.5) there exists a diagonal matrix $\hat{D}_{k-1}^*$ whose diagonal elements are the nonzero eigenvalues of $\widehat{\Omega}_m^*$ (in decreasing order of magnitude), a $k \times (k-1)$ matrix $\hat{P}_{k-1}^*$ whose columns are the (orthogonal) eigenvectors corresponding to the nonzero roots of $\widehat{\Omega}_m^*$, such that

$$\widehat{\Omega}_m^* = \hat{P}_{k-1}^* \hat{D}_{k-1}^* \hat{P}_{k-1}^{*\prime} \text{ and } \widehat{\Omega}_m^{*-} = \hat{P}_{k-1}^* \hat{D}_{k-1}^{*-1} \hat{P}_{k-1}^{*\prime}, \tag{A.10}$$

where $\hat{P}_{k-1}^*$ and $\hat{D}_{k-1}^*$ satisfy the following convergence:

$$\hat{P}_{k-1}^* \mid \mathscr{D}_n \xrightarrow{p} P_{k-1} \; a.s.. \text{ and } \hat{D}_{k-1}^* \mid \mathscr{D}_n \xrightarrow{p} D_{k-1} \; a.s., \tag{A.11}$$

where P_{k-1} and D_{k-1} are the matrices defined in equation (A.5) [in the proof of Lemma 2]. Now, from the proof of Lemma 3, we can express $\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ as:

$$\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \;\; = \;\; S_m^{*\prime} \widehat{\Omega}_m^{*-} S_m^* = \widetilde{S}_m^{*\prime} \widetilde{S}_m^*, \tag{A.12}$$

where $\widetilde{S}_m^* = \hat{D}_{k-1}^{*-1/2} \hat{P}_{k-1}^{*\prime} S_m^* = \sum_{i=1}^n \widetilde{R}_{m,i}^*$ and $\{\widetilde{R}_{m,i}^*\}_{i=1}^n$ are also i.i.d under $\mathbb{P}^*$. By adapting the proof of the Liapunov Central Limit Theorem in Lemma 3, we have

$$\widetilde{S}_m^* \mid \mathscr{D}_n \;\; \xrightarrow{d} \;\; N(0,\, I_{k-1}) \; a.s. \tag{A.13}$$

Moreover, since $\{\widetilde{R}_{m,i}^*\}_{i=1}^n$ are i.i.d under $\mathbb{P}^*$ with finite second moments, from the Berry-Esseen theorem for sums of independent random vectors, we have

$$\sup_{x \in \mathbb{R}^{k-1}} \left| \mathbb{P}^*(\widetilde{S}_m^* \leqslant x) - \Phi(x) \right| \leqslant \frac{c(k)}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}^*[\|\widetilde{R}_{m,i}^*\|^{2+\delta}], \tag{A.14}$$

where $c(k)$ is a constant that depends on k ($=$ dimension of $\widetilde{S}_m^*$), $\Phi(\cdot) \equiv$ cdf of $N(0,\, I_{k-1})$. Moreover, by adapting the proof of the Liapunov Central Limit Theorem in step (c) of the proof of Lemma 3, we have

$$\sum_{i=1}^{n} \mathbb{E}^*[\|\widetilde{R}_{m,i}^*\|^{2+\delta}] = \sum_{i=1}^{n} \mathbb{E}^*\Big[\|\hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} R_{m,i}^*\|^{2+\delta}\Big] \leqslant cn^{-\frac{\delta}{2}} n^{-1} \sum_{i=1}^{n} \mathbb{E}^*\Big[\|\hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (d_{g\cdot i}^* - d_{g'\cdot i}^*)\|^{2+}$$
$$+\|\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (u_{g\cdot i} - u_{g'\cdot i})\|^{2+\delta}\Big] = cn^{-\frac{\delta}{2}} \mathbb{E}^*\Big[\|\hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (d_{g\cdot i}^* - d_{g'\cdot i}^*)\|^{2+\delta} +$$
$$cn^{-\frac{\delta}{2}} \Big\|\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (u_{g\cdot i} - u_{g'\cdot i})\Big\|^{2+\delta} \qquad \text{(A.1}$$

for a large enough constant $c \in \mathbb{R}^+$. However, the second term of the last equality in (A.15) is such that:

$$cn^{-\frac{\delta}{2}} \Big\|\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (u_{g\cdot i} - u_{g'\cdot i})\Big\|^{2+\delta} \xrightarrow{p} 0 \; a.s.$$

because $\Big\|\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (u_{g\cdot i} - u_{g'\cdot i})\Big\|^{2+\delta} \xrightarrow{a.s.} \Big\|D_{k-1}^{-1/2} P_{k-1}' v_m\Big\|^{2+\delta} = 0$ under H_{0m} and $cn^{-\frac{\delta}{2}} \to 0$ as $n \to \infty$. Similarly, the first term of the last equality in (A.15) is such that:

$$cn^{-\frac{\delta}{2}} \mathbb{E}^*\Big[\|\hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (d_{g\cdot i}^* - d_{g'\cdot i}^*)\|^{2+\delta}\Big] \xrightarrow{p} 0 \; a.s.$$

because $\mathbb{E}^*\Big[\|\hat{D}_{k-1}^{*^{-1/2}} \hat{P}_{k-1}^{*'} \mathbf{T}^m (d_{g\cdot i}^* - d_{g'\cdot i}^*)\|^{2+\delta}\Big] \xrightarrow{p^*} \Big\|\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{k-1}^{-1/2} \hat{P}_{k-1}' \mathbf{T}^m (u_{g\cdot i} - u_{g'\cdot i})\Big\|^{2+\delta}$ and $\Big\|\frac{1}{n} \sum_{i=1}^{n} \hat{D}_{k-1}^{-1/2} \hat{P}_{k-1}' \mathbf{T}^m (u_{g\cdot i} - u_{g'\cdot i})\Big\|^{2+\delta} \xrightarrow{p} \Big\|D_{k-1}^{-1/2} P_{k-1}' v_m\Big\|^{2+\delta} = 0$ under H_{0m}; and in addition $cn^{-\frac{\delta}{2}} \to 0$ as $n \to \infty$. Therefore, we have $\sum_{i=1}^{n} \mathbb{E}^*[\|\widetilde{R}_{m,i}^*\|^{2+\delta}] \mid \mathscr{D}_n \xrightarrow{p} 0 \; a.s.$ in prob-$\mathbb{P}$, which entails that $\frac{c(k)}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}^*[\|\widetilde{R}_{m,i}^*\|^{2+\delta}] \mid \mathscr{D}_n \xrightarrow{p} 0 \; a.s.$ in prob-$\mathbb{P}$. From (A.14), it is clear that we have

$$\sup_{x \in \mathbb{R}^{k-1}} \Big|\mathbb{P}^*(\widetilde{S}_m^* \leqslant x) - \Phi(x)\Big| \xrightarrow{p} 0 \text{ in prob-}\mathbb{P}. \tag{A.16}$$

Now, by using (A.12), we can write $\mathbb{P}^*(\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w)$ as: $\mathbb{P}^*(\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w) = \mathbb{P}^*(\widetilde{S}_m^* \in \mathscr{C}_w)$ where $\mathscr{C}_w = \{x \in \mathbb{R}^{k-1} : \; x'x \leqslant w\}$ are convex sets in $\mathbb{R}^{k-1}$. From Bhattacharya and Rao (1976, Corollary 3.2), we have

$\sup\limits_{w\in\mathbb{R}}\Phi\big((\partial\mathscr{C}_w)^\epsilon\big) \leqslant d.\epsilon$ for some constant d and $\epsilon > 0$. Hence, Bhattacharya and Ghosh (1978, Theorem 1) holds with $W_n \equiv \mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I}))$ and $B \equiv \mathscr{C}_w$, thus

$$\sup_{w\in\mathbb{R}}\left|\mathbb{P}^*(\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w) - G_{k-1}(w)\right| \xrightarrow{p} 0 \text{ in prob-}\mathbb{P}, \tag{A.17}$$

where $G_{k-1}(\cdot) \equiv$ cdf of $\chi^2(k-1)$. Finally, we have: $\sup\limits_{w\in\mathbb{R}}\left|\mathbb{P}^*(\mathbf{W}_m^*(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w) - \mathbb{P}(\mathbf{W}_m(\mathbf{P}_n^{(k)}(\mathbf{I})) \leqslant w)\right| \xrightarrow{p} 0$ in prob-$\mathbb{P}$ by Lemma 2.

(b) Under H_{1m} or H_{2m}, the results follow straightforwardly from Lemma 2-(b) and Lemma 3-(b), so the proof is omitted.

Aliprantis, D. and Richter, F. G.-C. (2013). Evidence of neighborhood effects from mto: Lates of neighborhood quality. *Review of Economics and Statistics*, 88(3):389–432.

Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica*, 64(5):pp. 1183–1193.

Andrews, D. W. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica*, 70(1):119–162.

Atkinson, A. B. (1987). On the measurement of poverty. *Econometrica: Journal of the Econometric Society*, 55(4):749–764.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2012). The diffusion of microfinance. NBER Working Papers 17743, National Bureau of Economic Research, Inc.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144):1236498.

Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104.

Barrett, G. F., Donald, S. G., and Bhattacharya, D. (2014). Consistent nonparametric tests for lorenz dominance. *Journal of Business & Economic Statistics*, 32(1):1–13.

Beach, C. M. and Richmond, J. (1985). Joint confidence intervals for income shares and lorenz curves. *International Economic Review*, 26:439–450.

Bethge, M. and Berens, P. (2007). Near-maximum entropy models for binary neural representations of natural images. In *NIPS*, pages 97–104.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal edgeworth expansion. *The Annals of Statistics*, 6(2):434–451.

Bhattacharya, R. N. and Rao, R. R. (1976). *Normal approximation and asymptotic expansions*. John Wiley & Sons, Inc.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Bickel, P. J., Chen, A., Levina, E., et al. (2011). The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301.

Blume, L. E., Brock, W. A., Durlauf, S. N., and Jayaraman, R. (2015). Linear social interactions models. *Journal of Political Economy*, 123(2):444–496.

Borghans, L., Heckman, J. J., Golsteyn, B. H., and Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3):649–658.

Bramoulle, Y. and Kranton, R. (2007). Risk-sharing networks. *Journal of Economic Behavior & Organization*, 64(3-4):275–294.

Brown, B. W. and Newey, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business & Economic Statistics*, 20(4):507–517.

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273–1301.

Calv-Armengol, A. (2004). Job contact networks. *Journal of Economic Theory*, 115(1):191 – 206.

Calvo-Armengol, A. (2004). Job contact networks. *Journal of Economic Theory*, 115(1):191–206.

Chandrasekhar, A. (2015). Econometrics of network formation. *Oxford Handbook on the Econometrics of Networks*, 12.

Choi, S., Masson, V., Moore, A., and Oak, M. (2013). Networks and Favor Exchange Norms under Stochastic Costs. School of Economics Working Papers 2013-04, University of Adelaide, School of Economics.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345.

Davidson, R. and Duclos, J.-Y. (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica*, 68(6):1435–1464.

Dhrymes, P. J. (1978). *Mathematics for econometrics*. Springer.

Edwards, A. L. (1950). On" the use and misuse of the chi-square test"the case of the 2x2 contingency table. *Psychological bulletin*, 47(4):341.

Fafchamps, M. and Lund, S. (2003). Risk-sharing networks in rural philippines. *Journal of development Economics*, 71(2):261–287.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Goyal, S. (2012). *Connections: an introduction to the economics of networks*. Princeton University Press.

Graham, B. S. (2014). Methods of identification in social networks. Technical report, National Bureau of Economic Research.

Gumbel, E. J. (1943). On the reliability of the classical chi-square test. *The Annals of Mathematical Statistics*, 14(3):253–263.

Hahn, J. (1996). A note on bootstrapping generalized method of moments estimators. *Econometric Theory*, 12(01):187–197.

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica: Journal of the Econometric Society*, 64(4):891–916.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 46(6):1251–1271.

Hotelling, H. (1930). The consistency and ultimate distribution of optimum statistics. *Transactions of the American Mathematical Society*, 32(4):847–859.

Hsieh, C.-S. and Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2):301–319.

Inoue, A. and Shintani, M. (2006). Bootstrapping gmm estimators for time series. *Journal of Econometrics*, 133(2):531–555.

Jackson, M. O. (2014). Networks in the understanding of economic behaviors. *The Journal of Economic Perspectives*, 28(4):3–22.

Jackson, M. O. et al. (2008). *Social and economic networks*. Princeton University Press.

Jackson, M. O., Rodriguez-Barraquer, T., and Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, 102(5):1857–97.

Leung, M. P. (2015). Two-step estimation of network-formation models with incomplete information. *Journal of Econometrics*, 188(1):182–195.

Lewis, D. and Burke, C. J. (1949). The use and misuse of the chi-square test. *Psychological Bulletin*, 46(6):433.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765.

Liu, X. (2013). Estimation of a local-aggregate network model with sampled networks. *Economics Letters*, 118(1):243–246.

Macke, J. H., Berens, P., Ecker, A. S., Tolias, A. S., and Bethge, M. (2009). Generating spike trains with specified correlation coefficients. *Neural computation*, 21(2):397–423.

Mann, H. B. and Wald, A. (1942). On the choice of the number of class intervals in the application of the chi square test. *The Annals of Mathematical Statistics*, 13(3):306–317.

McFadden, D. (1989). Testing for stochastic dominance. In *Studies in the Economics of Uncertainty*, pages 113–134. Springer.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Jornal of Science*, 50(302):157–175.

Richmond, J. (1982). A general method for constructing simultaneous confidence intervals. *Journal of the American Statistical Association*, 77(378):455–460.

Rosenzweig, M. R. and Stark, O. (1989). Consumption smoothing, migration, and marriage: Evidence from rural india. *Journal of Political Economy*, 97(4):905–926.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. a definition. *Journal of Economic Theory*, 2(3):225 – 243.

Schorr, B. (1974). On the choice of the class intervals in the application of the chi-square test. *Statistics: A Journal of Theoretical and Applied Statistics*, 5(4-5):357–377.

Stoline, M. R. and Ury, H. K. (1979). Tables of the studentized maximum modulus distribution and an application to multiple comparisons among means. *Technometrics*, 21(1):pp. 87–93.

Tesfatsion, L. (1997). A trade network game with endogenous partner selection. In *Computational approaches to economic problems*, pages 249–269. Springer.

Williams, C. A. (1950). The choice of the number and width of classes for the chi-square test of goodness of fit. *Journal of the American Statistical Association*, 45(249):77–86.

Yates, F. (1934). Contingency table involving small numbers and the chi-square test. *Journal of the Royal Statistical Society*, 1(2):217–235.

Statement of Authorship

Title of Paper	**Inference for the Degree Distribution of a Graph**	
Publication Status	□ Published □ Submitted for Publication	□ Accepted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details		

Principal Author

Name of Principal Author (Candidate)	Robert Garrard		
Contribution to the Paper			
Overall percentage (%)	100%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	01/06/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Please cut and paste additional co-author panels here as required.

Inference for the Degree Distribution of a Graph

Robert Garrard

School of Economics, University of Adelaide

Abstract

We consider the problem of testing a hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}$ is a vector representing the degree distribution of a graph and the sample acquired is an induced subgraph. We propose a novel bootstrap procedure to control the size of a test under the null hypothesis by constructing a graph whose degree distribution conforms to the null hypothesis from which we may draw pseudo-samples in the form of induced subgraphs. We investigate the properties of the bootstrap with a simulation study in which a Wald-type statistic based on a truncated singular value estimator, whose null distribution is approximately chi-square, serves as a benchmark.

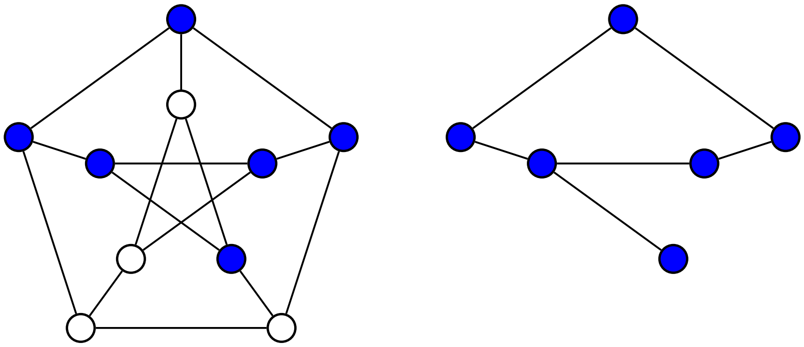
Keywords: Networks, degree distribution, induced subgraph, singular value decomposition, bootstrap
JEL: C12, C15, C45, D85

1. Introduction

Graphs are an increasingly popular tool used to model complex relationships and interactions, such as the spread of contagion through a financial system (Nier et al. 2007, Gai and Kapadia 2010), interhousehold risk sharing behavior (Bramoulle and Kranton 2007), trade in the absence of markets (Kranton and Minehart 2001), and even the interaction between proteins within a cell (Pellegrini et al. 2004).[1] One of the many salient features of a graph is its degree distribution, which captures the pattern of direct connections between nodes. Albert et al. (2000) show that the degree distribution is strongly tied to the ability of a graph to withstand failures of some of

Email address: `robert.garrard@adelaide.edu.au`

[1] The terms graph and network may be used interchangeably.



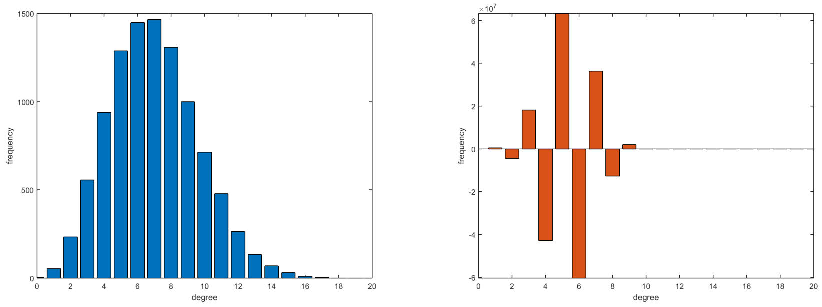
(a) Parent Graph (b) Induced Subgraph

Figure 1: 1a shows a parent graph representing a population of ten nodes where six nodes have been selected through simple random sampling. 1b displays the subgraph induced by those nodes.

its elements. Doyle et al. (2005) coined the term "robust-yet-fragile" to refer to graphs that are robust to random failures but vulnerable to targeted attacks or failures of certain key elements. This feature is common to many real-world graphs such as the webgraph of the internet or interbank lending networks in a financial system (Boss et al. 2004, Gai et al. 2011). Galeotti et al. (2010) show how equilibrium outcomes of games on networks are sensitive to changes in the degree distribution. Thus the ability to conduct inference regarding the degree distribution of a graph is paramount to understanding key aspects of real-world graphs.

Since graphs are often too large to observe in their entirety, inferences about their features must be made from sampled subgraphs. One common sampling method, induced subgraph sampling, involves taking a random sample of nodes and observing only the connections between those nodes sampled. This yields a sampled subgraph for which we may compute features of interest, such as measures of centrality, clustering, etc. However, this sampling design distorts the degree distribution by systematically ignoring links to nodes not in the sample. Specifically it maps a vector representing the degree distribution into a lower-dimension subspace in such a way that is difficult to invert. This may be modelled as a linear regression problem in which the regressors exhibit near multicollinearity. Figure 1 illustrates a sampled induced subgraph.

In this paper we consider how to test a simple hypothesis regarding the degree distribution of a graph upon observing an induced subgraph sam-



(a) True Distribution

(b) Unbiased Estimator

Figure 2: Estimate of degree distribution using an unbiased estimator. Sample of size $n = 2000$ drawn from a Poisson random graph on $N = 10,000$ nodes with probability parameter such that $Np = 7$.

ple. Our contribution is to propose a novel bootstrap procedure to control the size of a test under the null hypothesis. This procedure exploits the algorithm of the "configurations model" (Bender and Canfield, 1978; Bollobás, 1982) to construct a graph whose degree distribution conforms to the null hypothesis from which we may draw pseudo-samples in the form of induced subgraphs. Bhattacharyya and Bickel (2015) develop a subsampling procedure for estimating count features of a graph, but to the best of our knowledge this is the first paper to propose a bootstrap procedure in a setting which does not assume a form for the data generating process governing the formation of the population graph. However, this bootstrap procedure may only be applied to testing hypotheses regarding the degree distribution and not other count features of a graph. We investigate the performance of this procedure for a Wald-type statistic with a simulation study. Under appropriate conditions the null distribution of the statistic may be considered approximately chi-square which we use as a benchmark.

Construction of the test statistic requires choosing an estimator for the degree distribution. Frank (1980) was the first to approch this problem for which he constructed an unbiased estimator. However, this estimator tends to have very large mean square error. In particular, this estimator does not meet the basic requirement that elements of the degree distribution be between 0 and 1 and exhibits erratic sign switching behavior, as is illustrated in Figure 2.

Zhang et al. (2015) propose an estimator which requires that the estimate be a valid probability mass by constraint and overcomes the multicollinearity of the regressors with a smoothing penalty whose tuning parameter is chosen through Monte Carlo SURE (Ramani et al., 2008; Eldar, 2009). While this strategy enforces desirable behavior of the estimator, the data-driven choice of the tuning parameter makes the distribution of any statistic based on this estimator difficult to characterize. Additionally, the requirement that both the tuning parameter and the estimator itself be determined numerically are likely to make simulation studies of its properties burdensome on standard computers.

We propose estimating the degree distribution by truncating the singular value decomposition (TSVD) of the design matrix (Golub and Kahan, 1965); that is, retaining only the singular values which are sufficiently large. This results in an estimator whose variance is significantly smaller than that of Frank (1980) at the expense of becoming biased. The bandwidth parameter, which governs how many singular values are retained, is chosen non-randomly; namely to minimize estimation risk under the null hypothesis.

We conduct a simulation study to investigate the properties of this procedure under the null (size) and the alternative (power) on a Poisson random graph with 10,000 nodes. We find that for small sample sizes the test based on the chi-square approximation becomes significantly size distorted. The bootstrap procedure returns the size of the test close to the nominal level. We find that the bootstrapped test corresponding to the minimum risk estimator enjoys up to 35 percentage points more power than the test corresponding to the unbiased estimator.

The rest of the paper is organized as follows. Section 2 describes the construction of the test statistic and characterizes its distribution. Section 3 describes the bootstrap procedure and how to construct a graph conforming to the null hypothesis. Section 4 conducts a simulation study to determine the small sample properties of the procedure. Section 5 concludes with a discussion.

2. Setup

Consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{X}$ is an is an $n \times p$ matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is vector of observations, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of parameters, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of noise. This model may be viewed as representing how the degree distribution of an induced subgraph, $\boldsymbol{y}$, is on average a linear transformation of the degree distribution of the population graph, $\boldsymbol{\beta}$. The design matrix $\boldsymbol{X}$ is a non-random matrix whose elements govern how the degree distribution is distorted as a function of the number of nodes in the population graph and the number of nodes sampled.

The appearance of linear regression models in network analysis is not uncommon, especially in the literature on the identification of peer effects (Manski, 1993; Bertrand et al., 2000; Gaviria and Raphael, 2001; Bramoull et al., 2009; Goldsmith-Pinkham and Imbens, 2013). In that setting, $\boldsymbol{y}$ is a vector of outcome variables and $\boldsymbol{X}$ is a matrix whose columns are random covariates, a subset of which describe the peers each agent is linked to in a network. The objective is to identify the effect that attributes of an agent's peers have on that agent's outcome. This typically involves endogeneity of the peer effect, since the network is formed by agents mutually agreeing to share a link, which must be resolved either through intrumental variable methods or structural modelling of the network formation process. Our setting, on the other hand, is more akin to the literatures on compressed sensing (Donoho, 2006; Candes and Tao, 2007) in which $\boldsymbol{X}$ is a matrix of non-random elements representing some sort of sensing apparatus which is attempting to measure $\boldsymbol{\beta}$ and in doing so maps $\boldsymbol{\beta}$ into a lower-dimensional space, and the literature on ill-posed inverse problems (O'Sullivan, 1986; Hansen, 1998), in which recovery of $\boldsymbol{\beta}$ is attempted when $\boldsymbol{X}$ experiences extreme multicollinearity. However, following Frank (1980) and Zhang et al. (2015), it is often convenient to assume that there is a known largest degree in the population graph which allows us to truncate the rows and columns of $\boldsymbol{X}$ such that it becomes square.

We are concerned with making inferences on $\boldsymbol{\beta}$ in (1). In particular, we consider testing the hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{against} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \tag{2}$$

where $\boldsymbol{\beta}_0$ is a hypothesised degree distribution. We begin first with the construction of an estimator for $\boldsymbol{\beta}$. Section 2.1 will examine the distribution of the sampling error $\boldsymbol{\varepsilon}$, section 2.2 will construct a statistic against which we will benchmark the bootstrap, and section 2.3 will discuss the choice of tuning parameter corresponding to the estimator. Section 5 discusses the inversion of this test to construct confidence intervals.

Using the well known singular value decomposition, $\boldsymbol{X}$ may be written as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}' \tag{3}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times n$ and $p \times p$ orthogonal matrices respectively, and $\boldsymbol{\Sigma}$ is an $n \times p$ rectangular diagonal matrix whose elements, $\sigma_1 \geq \cdots \geq \sigma_{\min\{n,p\}}$, are called singular values. The condition number of a matrix, κ, is the ratio of the largest singular value to the smallest.

$$\kappa(\boldsymbol{X}) = \frac{\sigma_{\max}(\boldsymbol{X})}{\sigma_{\min}(\boldsymbol{X})} \tag{4}$$

Matrices for which $\sigma_{\min} = 0$ are called singular and by convention have condition number $\kappa(\boldsymbol{X}) = \infty$. Matrices with a large condition number are said to be ill-conditioned. In econometrics this is typically caused by high multicollinearity among the columns of the design matrix. When all singular values are non-zero, the standard Moore-Penrose pseudo-inverse may be constructed by inverting the singular value decomposition.

$$\boldsymbol{X}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}' \tag{5}$$

where $\boldsymbol{\Sigma}^{-1}$ denotes a $p \times n$ rectangular diagonal matrix whose elements are the reciprocal of each singular value. From this we may construct an estimator for $\boldsymbol{\beta}$ in (1).

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{X}^+\boldsymbol{y} \tag{6}$$

In the low-dimensional setting, where $n > p$, $\tilde{\boldsymbol{\beta}}$ returns the usual Ordinary Least Squares (OLS) estimator which uniquely minimizes the sum of squared residuals. In the high-dimensional setting ($p > n$), a continuum of solutions for $\boldsymbol{\beta}$ exist which yield perfect fit (the residual vector is zero). In this case $\tilde{\boldsymbol{\beta}}$ returns the solution with minimum ℓ_2-norm.

Construction of the Moore-Penrose pseudo-inverse requires taking the reciprocal of each singular value. For ill-conditioned matrices, which have singular values tending toward zero, the estimator in (6) can be highly unstable. When $\boldsymbol{y}$ is measured with even small amounts of noise, the value of the estimator may exhibit erratic sign switching behavior and coefficients may differ from their true values by several orders of magnitude, recall the unbiased estimator in Figure 2.

We may attempt to overcome this instability by inverting the singular value decomposition of $\boldsymbol{X}$ in such a way that the reciprocals of singular values which are "too small" are thresholded to zero. We may construct an estimator based on the truncated singular value decomposition (TSVD) due to Golub and Kahan (1965) as follows.

$$\boldsymbol{X}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{U}' \tag{7}$$

where $\boldsymbol{\Sigma}^{\dagger}$ is a $p \times n$ rectangular diagonal matrix with

$$\boldsymbol{\Sigma}^{\dagger}_{ii} = \begin{cases} \frac{1}{\sigma_i} & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Hence we may construct an estimator for $\boldsymbol{\beta}$ analogous to (6) with

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\dagger}\boldsymbol{y} \tag{9}$$

Setting $k = \min\{n, p\}$ thresholds no singular values and retrieves the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$. Setting $k < \min\{n, p\}$ gives an estimator wherein the variance of the estimator is reduced, yielding smoother estimates, at the expense of increased bias. We address the selection of the tuning parameter k in subsection 2.3 after characterizing an approximation for the distribution of $\boldsymbol{\varepsilon}$.

2.1. The Degree Distribution of an Induced Subgraph

Let $G = (V, E)$ be a graph and let $V' \subseteq V$ and $E' \subseteq E$ be a subset of nodes and edges. $G' = (V', E')$ is said to be an *induced subgraph* when any pair of nodes $a, b \in V'$ are adjacent in G' if and only if they are adjacent in G.[2] An *induced subgraph sample* is formed by simple random sampling (SRS) a subset of nodes and constructing the subgraph induced by those nodes.

Suppose we draw an induced subgraph sample G' from a graph G. Let $N = |V|$ and $n = |V'|$ be the number of nodes in the population and sample respectively, where $|\cdot|$ denotes cardinality. Let $\boldsymbol{\beta} \in \mathbb{R}^N$ represent the degree distribution of G such that $\boldsymbol{\beta}_i$ is the proportion of nodes in G with degree $i = 0, \ldots, N-1$, and let $\boldsymbol{y} \in \mathbb{R}^n$ be defined similarly for G'.

Proposition 1. *Let n, N, $\boldsymbol{y}$, and $\boldsymbol{\beta}$ be defined as above. Then*

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta} \tag{10}$$

[2] For further reference see Wilson (1996).

where

$$\boldsymbol{X}_{ij} = \binom{j}{i}\binom{N-1-j}{n-1-i}\binom{N-1}{n-1}^{-1} \tag{11}$$

for i, j such that the binomial coefficients above are well defined and $X_{ij} = 0$ otherwise.

Proof. Consider a node $i \in G'$. Let d_i and d_i' denote the degree of i in G and G' respectively. For $0 \leq x \leq n-1$ and $0 \leq y \leq N-1$, consider $\mathbb{P}(d_i' = x \mid d_i = y)$. Conditional on i having y edges in the population graph, to have x links in the subgraph induced on i and the remaining $n-1$ nodes requires x successes out of a possible y successes from a sample of $n-1$ drawn without replacement. Thus the conditional probability that i has x links in the subgraph is hypergeometric.

$$\mathbb{P}(d_i' = x \mid d_i = y) = \frac{\binom{y}{x}\binom{N-1-y}{n-1-x}}{\binom{N-1}{n-1}} \tag{12}$$

Marginalizing over d_i yields (10) and (11). □

Thus we may model the degree distribution of an induced subgraph under SRS of nodes with the linear model in (1).

It is noteworthy that this sampling design yields a rarity in applied statistics, namely that the true regression function is known a priori to be linear and that the design matrix is non-random. This typically only occurs as a specially designed feature of a sensing apparatus, such as in medical imaging.

As noted in Zhang et al. (2015), the design matrix in subgraph sampling is severely ill-conditioned. Figure 3 shows a scree plot of the first few singular values of a design matrix based on sampling 10% of nodes in a population of 10,000. We see that the singular values decay exponentially to zero, leading to instability of the Moore-Penrose based estimator. Figure 4 shows the improved performance in estimating the degree distribution from the same sample, but using only the first 3 singular values.

While the TSVD estimator yields significant improvement over the unbiased estimator of Frank (1980), in order to conduct inference we first need to characterize the distribution of the sampling error, $\boldsymbol{\varepsilon}$. We have the following result from Zhang et al. (2015).

Proposition 2 (Zhang et al.)**.** *For large sample size, n, and small sampling rate, $\frac{n}{N}$, the sampling error may be approximated by $\boldsymbol{\varepsilon} \sim N(0, \frac{1}{n} diag(\boldsymbol{X\beta}))$.*

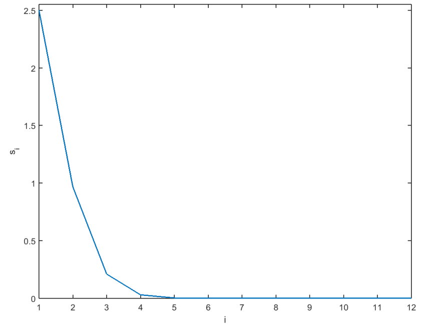
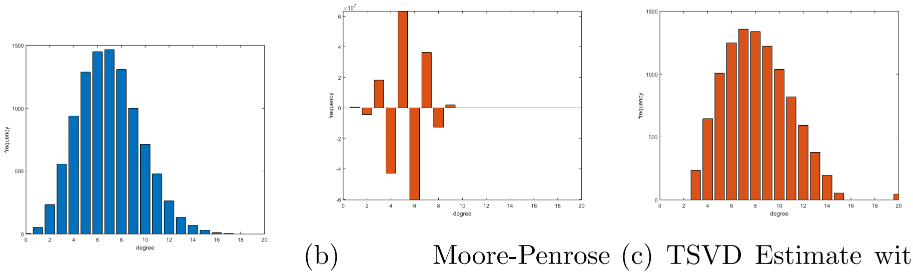


Figure 3: Exponential decay of singular values for a design matrix corresponding to induced subgraph sampling of $n = 1,000$ nodes from a population of $N = 10,000$.



(a) True Distribution

(b) Moore-Penrose Inverse

(c) TSVD Estimate with k=3

Figure 4: TSVD estimate for degree distribution. Sample of $n = 1,000$ drawn from a Poisson random graph with $N = 10,000$ nodes and link probability such that $Np = 7$.

Proof. See Zhang et al. (2015, Sec 2.3 and App B). The scaling factor $\frac{1}{n}$ comes from the fact that we use proportions instead of counts. $\square$

Specifically, they show using the Chen-Stein method (Chen, 1975; Barbour et al., 1992) that for each i, the law of $\boldsymbol{y}_i$ is close in total variation distance to a Poisson distribution with intensity $\mu_i = \mathbb{E}[\boldsymbol{y}_i]$. For large n and small $\frac{n}{N}$ this may be well approximated by a normal distribution and off-diagonal entries of the covariance matrix are approximately zero.

2.2. Test statistic

Let $\boldsymbol{V}$, $\boldsymbol{\Sigma}$ be defined as in (3). Let $\mathcal{I} = \{i \mid (\boldsymbol{X}\boldsymbol{\beta}_0)_i \neq 0, i = 1, \ldots, k\}$, $\ell = |\mathcal{I}|$, $\tilde{\boldsymbol{V}}$ be the set of $\mathcal{I}$ columns of $\boldsymbol{V}$, $\boldsymbol{\Lambda} = n^{-1}\mathrm{diag}(\boldsymbol{X}\boldsymbol{\beta}_0)$, $\tilde{\boldsymbol{\Lambda}}$ be the $\mathcal{I}$ rows and columns of $\boldsymbol{\Lambda}$, $\tilde{\boldsymbol{\Sigma}}$ be the $\mathcal{I}$ rows and columns of $\boldsymbol{\Sigma}$, and $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{\Sigma}}^2\tilde{\boldsymbol{\Lambda}}^{-1}$. Consider the statistic

$$\boldsymbol{W}(\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}} - \boldsymbol{X}^\dagger \boldsymbol{X} \boldsymbol{\beta}_0)' \left[\tilde{\boldsymbol{V}} \tilde{\boldsymbol{\Omega}} \tilde{\boldsymbol{V}}' \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{X}^\dagger \boldsymbol{X} \boldsymbol{\beta}_0) \tag{13}$$

Suppose we assume that the approximation in proposition 2 holds exactly with $\boldsymbol{\varepsilon} \sim N(0, \frac{1}{n}\text{diag}(\boldsymbol{X}\boldsymbol{\beta}))$. Then we have that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{X}^\dagger \boldsymbol{X} \boldsymbol{\beta}_0 = \boldsymbol{X}^\dagger \boldsymbol{X} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{X}^\dagger \boldsymbol{\varepsilon} \sim N(\boldsymbol{X}^\dagger \boldsymbol{X} (\boldsymbol{\beta} - \boldsymbol{\beta}_0), \boldsymbol{X}^\dagger \boldsymbol{\Lambda} \boldsymbol{X}^{\dagger\prime}) \tag{14}$$

Thus with an appropriate scaling, a quadratic form in $\hat{\boldsymbol{\beta}} - \boldsymbol{X}^\dagger \boldsymbol{X} \boldsymbol{\beta}_0$ will have a χ^2 distribution.

Lemma 1. *Under the above definitions*

$$\tilde{\boldsymbol{\Omega}}^{1/2} \tilde{\boldsymbol{V}}' \left(\boldsymbol{X}^\dagger \boldsymbol{\Lambda} \boldsymbol{X}^{\dagger\prime} \right) \tilde{\boldsymbol{V}} \tilde{\boldsymbol{\Omega}}^{1/2} = \boldsymbol{I}_\ell \tag{15}$$

Proof is in the appendix. Therefore, under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ we have that

$$\boldsymbol{W}(\boldsymbol{\beta}_0) \sim \chi^2(\ell) \tag{16}$$

To serve as a benchmark we will use the statistic in (16) with critical values drawn from the appropriate χ^2 distribution. Note that since we are not directly estimating diag($\boldsymbol{X}\boldsymbol{\beta}$) but rather using its value *under the null*, when in the alternative hypothesis this test statistic does not attain the usual non-central chi-square distribution.

2.3. Selection of the tuning parameter

We now turn to the question of how many singular values should be retained when constructing the estimator. A natural choice would be to select k to minimize estimation risk.

$$R(k) = \mathbb{E}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 \tag{17}$$

where $||\cdot||_2$ denotes the ℓ_2-norm. However, the estimation risk depends on the unknown parameter $\boldsymbol{\beta}$. Exploiting the Gaussian approximation in proposition 2 we could form an unbiased estimator of the risk function using Stein's unbiased risk estimate (SURE), due to Stein (1981). Supposing we have the linear model in (1) with a linear smoother $\boldsymbol{S}$ such that $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and $\boldsymbol{\varepsilon} \sim N(0, s^2\boldsymbol{I})$, SURE gives the following unbiased estimate of risk.

$$\hat{R} = -ns^2 + ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2 + 2s^2 tr(\boldsymbol{S}) \tag{18}$$

However, our setting is heteroscedastic. Since $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X\beta}$, we may proceed by using the unbiased estimator $\hat{\boldsymbol{\Lambda}} = diag(\boldsymbol{y})$ and conducting the suitable reweighting to obtain $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{\varepsilon}} \sim N(0, \boldsymbol{I})$.[3] Noting that in our setting $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^{\dagger}$ and

$$tr(\boldsymbol{X}\boldsymbol{X}^{\dagger}) = tr(\boldsymbol{U\Sigma V'V\Sigma^{\dagger}U'}) = tr(\boldsymbol{U\Sigma\Sigma^{\dagger}U'}) = tr(\boldsymbol{U}\begin{bmatrix}\boldsymbol{I}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{U}')$$
$$= tr(\boldsymbol{U'U}\begin{bmatrix}\boldsymbol{I}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}) = tr(\begin{bmatrix}\boldsymbol{I}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}) = k$$

We may choose the tuning parameter to minimize the unbiased estimate of risk.

$$k = \underset{k'}{\text{argmin}}\ ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}(k')||_2^2 + 2k' \tag{19}$$

Note that this is similar to using an information criterion such as Mallows' C_p (Mallows, 1995) or AIC (Akaike, 1973, 1974).

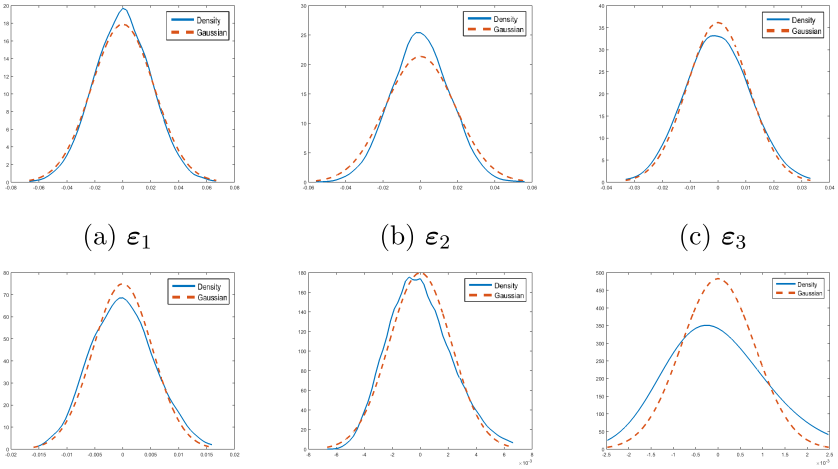
While this approach prescribes a way to select the tuning parameter such that the estimate of $\boldsymbol{\beta}$ has small risk, it happens to be data-driven, using the random variable $\boldsymbol{y}$ to assess goodness of fit. This results in k being chosen randomly, which makes the null distribution of a test statistic based on k difficult to characterize. If our goal were only to have an efficient point estimate for $\boldsymbol{\beta}$, this procedure would suffice. Since our goal is inference, we need a non-random way to choose k. While there are many options for such a procedure, we consider choosing k to minimize estimation risk *under the null*. Using the bias-variance decomposition of (17) we may write this as follows.

$$k = \underset{k'}{\text{argmin}}\quad tr\left(\boldsymbol{X}^{\dagger}\frac{1}{n}diag(\boldsymbol{X\beta}_0)\boldsymbol{X}^{\dagger\prime}\right) + ||(\boldsymbol{X}^{\dagger\prime}\boldsymbol{X} - \boldsymbol{I})\boldsymbol{\beta}_0||_2^2 \tag{20}$$

3. Bootstrap

For relatively small n or large $\frac{n}{N}$, the approximation of the distribution of noise to Gaussian may be quite poor, particularly for entries i where

[3] This is the standard GLS weighting except that some elements of $\boldsymbol{y}$ may be zero, causing rows of $\hat{\Lambda}$ being entirely zero and therefore $\hat{\Lambda}$ not positive definite. Let $\bar{\Lambda}$ be an $\ell \times n$ matrix formed by deleting the zero rows of $\hat{\Lambda}$. Then let $\tilde{y} = \bar{\Lambda}\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \bar{\Lambda}\boldsymbol{X}$, and proceed as above.



(a) $\boldsymbol{\varepsilon}_1$ (b) $\boldsymbol{\varepsilon}_2$ (c) $\boldsymbol{\varepsilon}_3$

(d) $\boldsymbol{\varepsilon}_4$ (e) $\boldsymbol{\varepsilon}_5$ (f) $\boldsymbol{\varepsilon}_6$

Figure 5: Distribution of sampling error compared to Gaussian approximation. 10,000 samples taken from a Poisson random graph with mean 7. We only display distributions for the first 6 entries of $\boldsymbol{\varepsilon}$ since nodes of degree greater than 5 have near 0 frequency in the induced subgraphs. Dashed lines represent the Gaussian approximation, solid lines the simulated distribution.

$(\boldsymbol{X\beta})_i$ is close to zero. Figure 5 compares the simulated distribution of noise for induced subgraphs drawn from a Poisson random graph to their respective Gaussian approximations for the first few terms of the support. This departure from normality may cause the hypothesis test to be size distorted. The typical correction under fixed design and heteroscedasticity is the wild bootstrap (Liu et al., 1988). However, this would appear not to be appropriate here as the covariance matrix has low rank. Bhattacharyya and Bickel (2015) propose a subsampling procedure for estimating variances for count features of graphs. This procedure involves resampling subgraphs at a much lower sampling rate. We propose a similar bootstrap-type procedure for testing the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ in which we construct a graph whose degree distribution conforms to $\boldsymbol{\beta}_0$ from which we draw pseudosamples of size n.

Suppose we had at our disposal a "null graph" $\mathcal{G}$ on N nodes whose degree distribution conformed to the null hypothesis $\boldsymbol{\beta}_0$. We could employ a typical parametric bootstrap to mimic the sampling distribution of $\hat{\boldsymbol{\beta}}$.

Algorithm 1 Graphical Bootstrap

```
$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\dagger}\boldsymbol{y}$, $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\beta}_0)$
for $i = 1$ to $B$ do
    $H \leftarrow$ induced subgraph of $\mathcal{G}$ from SRS of $n$ nodes
    $\boldsymbol{y}^* \leftarrow$ degree distribution of H.
    Construct bootstrap statistics $\hat{\boldsymbol{\beta}}_i^*$ and $\boldsymbol{W}_i^*$.
end for
$P \leftarrow \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{I}(\boldsymbol{W}_i^* > \boldsymbol{W})$
Reject $H_0$ if $P < \alpha$.
```

We next address the question of how to construct such a graph $\mathcal{G}$.

3.1. Sampling Graphs with a Prescribed Degree Distribution

Let $d_1, \ldots, d_N$ be the set of degrees of each node implied by the degree distribution. We wish to construct a graph with this degree sequence.[4] Begin with a set of nodes, $i = 1, \ldots, N$. Endow each node i, with a set of d_i *stubs* (or *half-links*) emanating from the node. Now construct a random matching on the set of stubs, and connect each pair of stubs that are matched to form a link.

To randomly match the stubs, create a list of the node labels in which label i appears d_i times, then form a random permutation of the list. To construct the graph, start at the first entry in the permuted list and begin pairing off the stubs of nodes with adjacent labels.

Example 1. Suppose we wish to construct a graph with degree sequence 4, 2, 2, 1, 1.

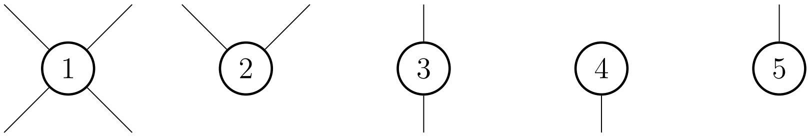


Figure 6: Nodes with 4, 2, 2, 1, and 1 stubs respectively.

Construct the list: 1111223345. Produce a random permutation: $\boxed{51}\boxed{13}\boxed{21}\boxed{23}\boxed{14}$. Connect adjacent nodes.

[4] The degree sequence is simply a list containing the degree of each node.

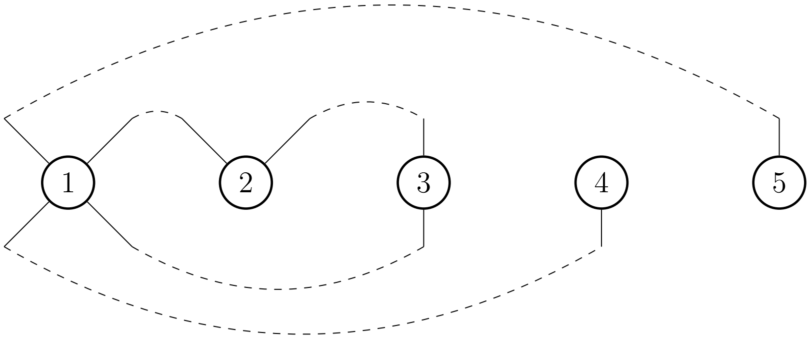


Figure 7: Connect the stubs.

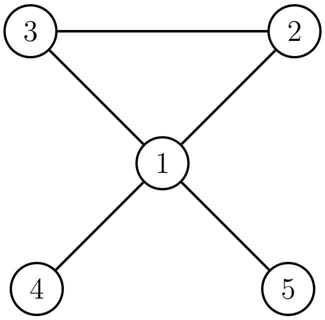


Figure 8: Resulting Graph

Given a degree sequence, $d_1, \ldots, d_N$, We construct the adjacency matrix of the null graph $\mathcal{G}$ on N nodes as follows.

Algorithm 2 Null Graph

```
Let $S$ be a vector of labels $1, \ldots, N$ where label $i$ has multiplicity $d_i$
$A \leftarrow zeros(N, N)$
$P \leftarrow randperm(S)$
for $i = 1$ to $N - 1$ Step 2 do
    $A(P(i), P(i + 1)) \leftarrow 1$
    $A(P(i + 1), P(i)) \leftarrow 1$
end for
```

This yields an adjacency matrix A from which we may sample induced subgraphs. The adjacency matrix for the subgraph induced by the set of nodes $J \subset \{j \mid j = 1, \ldots, N\}$ may be formed from extracting the set of J rows and corresponding columns from A.

Under this sampling method it is possible for the graph $\mathcal{G}$ to have self-loops and multiple edges between nodes. This may either be ignored, as the probability of such an occurance is declining with the number of nodes, or the algorithm may be repeated until a simple graph is formed.

4. Simulation Study

We investigate the properties of the hypothesis test under the null (size) and alternative (power) hypotheses with the following simulation study. Using algorithm 2, we construct adjacency matrices for population graphs on 10,000 nodes whose degree distribution is Poisson with means $\lambda = \{3, 3.5, 4, \ldots, 7\}$. For each λ we test the null hypothesis, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ against the alternative $H_A : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}_0$ is calculated from a Poisson distribution with rate parameter $\mu = 5$ at the 5% nominal level. Thus $\lambda = 5$ corresponds to the null hypothesis, while $\lambda \neq 5$ corresponds to the alternative. This is done for sampling rates $r \in \{2\%, 5\%, 10\%, 15\%, 20\%, 30\%\}$. The design matrix $\boldsymbol{X}$ is constructed according to (11) for $i, j \in 0, \ldots, 18$.[5] For each sampling rate and each λ we calculate the rejection frequency for the test of the null for the following procedures: (1) k is chosen to minimize risk under the null with critical values drawn according to equation (19); (2) the same choice of k with bootstrapped p-values; and (3) $k = 19$, corresponding to the Wald test based on the unbiased estimator. Table 1 displays the simulation results where each rejection frequency is calculated using 1,000 iterations and $B = 200$ bootstrap pseudo-samples.

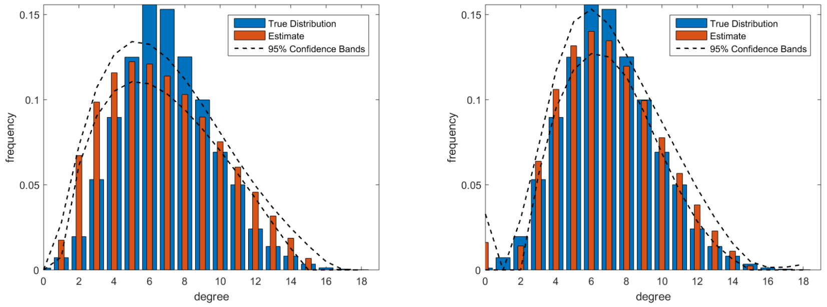
Under the null ($\lambda = 5$) we can see that the regularized test experiences significant size distortion, especially in smaller sample sizes. The bootstrap procedure appears to correct for this quite well with rejection rates close to the 5% nominal level. With regard to power, we see that the unbiased approach is overly conservative, even under the null. The bootstrap procedure offers a substantial improvement in power, offering around 35 percentage points more power under the alternative $\lambda = 3$ at the 5% sampling rate than the procedure based on an unbiased estimator.

[5] Since a Poisson distribution with mean 7 has almost zero mass outside this range, we may truncate the rows and columns of $\boldsymbol{X}$ for improved speed.

		λ								
Sampling Rate	Procedure	3.00	3.50	4.00	4.50	5.00	5.50	6.00	6.50	7.00
2%	(1)	33.8	21.0	14.3	10.7	9.5	11.5	18.3	26.3	33.8
	(2)	23.6	13.4	9.5	6.9	6.5	8.3	12.7	20.6	26.6
	(3)	0.9	1.3	1.3	2.3	3.4	6.2	8.7	14.0	18.1
5%	(1)	94.2	74.1	40.1	19.4	14.5	22.0	43.2	67.0	84.2
	(2)	83.8	55.3	23.1	8.8	4.9	10.2	26.2	46.7	69.9
	(3)	49.1	19.0	4.0	1.5	1.6	6.6	18.2	38.9	63.6
10%	(1)	100.0	99.7	88.4	35.2	12.3	32.5	79.9	98.6	100.0
	(2)	100.0	99.0	77.4	20.7	5.4	21.2	66.7	95.7	100.0
	(3)	100.0	97.1	44.0	5.7	2.8	13.4	55.7	92.4	99.9
15%	(1)	100.0	100.0	99.2	54.0	7.5	50.3	98.5	100.0	100.0
	(2)	100.0	100.0	98.7	47.8	5.8	42.8	97.3	100.0	100.0
	(3)	100.0	100.0	94.7	18.2	3.2	32.4	95.7	100.0	100.0
20%	(1)	100.0	100.0	100.0	73.3	3.2	68.7	100.0	100.0	100.0
	(2)	100.0	100.0	100.0	77.2	5.0	74.6	100.0	100.0	100.0
	(3)	100.0	100.0	99.9	45.6	3.4	57.8	99.7	100.0	100.0
30%	(1)	100.0	100.0	100.0	96.6	12.9	94.2	100.0	100.0	100.0
	(2)	100.0	100.0	100.0	88.5	4.9	83.9	100.0	100.0	100.0
	(3)	100.0	100.0	100.0	95.5	3.7	97.7	100.0	100.0	100.0

Table 1: Rejection frequency

5. Discussion



(a) 20% Sampling Rate

(b) 40% Sampling Rate

Figure 9: Confidence bands based on TSVD estimator with tuning parameter chosen according to 19 and $\hat{\Lambda} = \text{diag}(\boldsymbol{y})$. Population graph is Poisson on $N = 10,000$ nodes with probability parameter such that $Np = 7$.

The natural progression would be to consider the dual problem of constructing a confidence region for the degree distribution. One approach would be to construct a density estimate, for example based on the truncated singular value decomposition used here together with the tuning parameter chosen according to (19). Confidence bands centered at this estimate may be constructed through estimating standard errors with $\hat{\Lambda} = \text{diag}(\boldsymbol{y})$. The estimator's bias is quite large for even moderate sampling rates and so this confidence region will not be appropriately centered, leading to coverage well below the advertized rate, as is illustrated in fig. 9. An avenue for further research is to determine if the bias of this estimator may be well approximated or reduced without inflating its variance.

An alternative approach might be to invert the hypothesis test proposed here and take the appropriate projection of the resulting ellipsoid. Since the design is ill-conditioned, many highly non-smooth distributions are likely to be present in this ellipsoid such that resulting confidence bands will be overly wide. This could be ameliorated supposing one could invert the test only for distributions with sufficient smoothness. This paper has been primarily concerned with testing a simple hypothesis, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. However, it is unlikely in an empirical setting that there will be a clear choice for $\boldsymbol{\beta}_0$. A direction for further research would be in conducting hypothesis tests regarding whether an induced subgraph sample came from a particular *family* of distributions. In particular, whether the population graph is a poisson random graph, with poisson degree distribution, or scale-free graph, whose degree distribution follows a power law. Each of these distributions have different implications regarding both fragility of the graph to failure of elements and to revealing the process by which the graph may have formed.

Appendix

Proof of Lemma 1

Substituting in the definitions of $\tilde{\mathbf{\Omega}}$ and $\boldsymbol{X}^{\dagger}$, the LHS of lemma 1 becomes

$$\tilde{\mathbf{\Lambda}}^{-1/2}\left(\tilde{\mathbf{\Sigma}}\tilde{\boldsymbol{V}}'\boldsymbol{V}\mathbf{\Sigma}^{\dagger}\right)\boldsymbol{U}'\mathbf{\Lambda}\boldsymbol{U}\left(\mathbf{\Sigma}^{\dagger\prime}\boldsymbol{V}'\tilde{\boldsymbol{V}}\tilde{\mathbf{\Sigma}}'\right)\tilde{\mathbf{\Lambda}}'^{-1/2} \tag{21}$$

Note that since $\mathbf{\Lambda}$ is diagonal and $\boldsymbol{U}$ is orthogonal, $\boldsymbol{U}'\mathbf{\Lambda}\boldsymbol{U} = \mathbf{\Lambda}$. Note also that $\tilde{\boldsymbol{V}}'\boldsymbol{V}$ is an $\ell \times p$ matrix where the $\mathcal{I}$ columns form an ℓ dimensional identity matrix and the remaining columns are zero. Thus $\tilde{\mathbf{\Sigma}}\tilde{\boldsymbol{V}}'\boldsymbol{V}\mathbf{\Sigma}^{\dagger}$ is an $\ell \times n$ matrix whose $\mathcal{I}$ columns form an identity matrix and the remaining columns are zero. Therefore

$$\left(\tilde{\mathbf{\Sigma}}\tilde{\boldsymbol{V}}'\boldsymbol{V}\mathbf{\Sigma}^{\dagger}\right)\mathbf{\Lambda}\left(\mathbf{\Sigma}^{\dagger\prime}\boldsymbol{V}'\tilde{\boldsymbol{V}}\tilde{\mathbf{\Sigma}}'\right) = \tilde{\mathbf{\Lambda}} \tag{22}$$

and eq. (21) becomes

$$\tilde{\mathbf{\Lambda}}^{-1/2}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^{-1/2} = \boldsymbol{I}_{\ell} \tag{23}$$

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *nature*, 406(6794):378–382.

Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson approximation*. Clarendon Press Oxford.

Bender, E. A. and Canfield, E. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296 – 307.

Bertrand, M., Luttmer, E. F., and Mullainathan, S. (2000). Network effects and welfare cultures. *The Quarterly Journal of Economics*, 115(3):1019–1055.

Bhattacharyya, S. and Bickel, P. J. (2015). Subsampling bootstrap of count features of networks. *Ann. Statist.*, 43(6):2384–2411.

Bollobás, B. (1982). The asymptotic number of unlabelled regular graphs. *J. London Math. Soc*, 26(2):201–206.

Boss, M., Elsinger, H., Summer, M., and 4, S. T. (2004). Network topology of the interbank market. *Quantitative Finance*, 4(6):677–684.

Bramoull, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41 – 55.

Bramoulle, Y. and Kranton, R. (2007). Risk-sharing networks. *Journal of Economic Behavior & Organization*, 64(3-4):275–294.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability*, 3(3):534–545.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R., and Willinger, W. (2005). The robust yet fragile nature of the internet. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14497–14502.

Eldar, Y. C. (2009). Generalized sure for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481.

Frank, O. (1980). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 40(1):45 – 50.

Gai, P., Haldane, A., and Kapadia, S. (2011). Complexity, concentration and contagion. *Journal of Monetary Economics*, 58(5):453 – 470. Carnegie-Rochester Conference on public policy: Normalizing Central Bank Practice in Light of the credit Turmoi, 1213 November 2010.

Gai, P. and Kapadia, S. (2010). Contagion in financial networks. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., and Yariv, L. (2010). Network games. *The Review of Economic Studies*, 77(1):218–244.

Gaviria, A. and Raphael, S. (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics*, 83(2):257–268.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Golub, G. and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 2(2):205–224.

Hansen, P. C. (1998). *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM.

Kranton, R. E. and Minehart, D. F. (2001). A Theory of Buyer-Seller Networks. *American Economic Review*, 91(3):485–508.

Liu, R. Y. et al. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708.

Mallows, C. L. (1995). More comments on cp. *Technometrics*, 37(4):362–372.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531.

Nier, E., Yang, J., Yorulmazer, T., and Alentorn, A. (2007). Network models and financial stability. *Journal of Economic Dynamics and Control*, 31(6):2033 – 2060. Tenth Workshop on Economic Heterogeneous Interacting AgentsWEHIA 2005.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518.

Pellegrini, M., Haynor, D., and Johnson, J. M. (2004). Protein interaction networks. *Expert Review of Proteomics*, 1(2):239–249.

Ramani, S., Blu, T., and Unser, M. (2008). Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Transactions on Image Processing*, 17(9):1540–1554.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Wilson, R. (1996). *Introduction to Graph Theory*. Longman Group LTD, 4th edition.

Zhang, Y., Kolaczyk, E. D., and Spencer, B. D. (2015). Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.*, 9(1):166–199.

Statement of Authorship

Title of Paper	**On Lasso Methods for Econometric Inference**	
Publication Status	□ Published □ Submitted for Publication	□ Accepted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details		

Principal Author

Name of Principal Author (Candidate)	Robert Garrard		
Contribution to the Paper			
Overall percentage (%)	100%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	01/06/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Please cut and paste additional co-author panels here as required.

On Lasso Methods for Econometric Inference

Robert Garrard

School of Economics, University of Adelaide

Abstract

This paper presents a selective review of the Lasso estimator as it applies to econometric inference. We survey key papers addressing properties of the Lasso of interest to the econometrician including conditions for consistency, the asymptotic distribution of the estimator, its ability to be bootstrapped, sample splitting for high dimensional inference, and how it may be used to solve the many instruments problem in instrumental variables regression.

1. Introduction

Statistical and machine learning has garnered much interest lately as a methodology well suited to the world of Big Data. In such a world, the number of features being measured is close to, or sometimes greater than, the total number of observations. This renders classical statistical methods such as Ordinary Least Squares (OLS) infeasible. To regain feasibility, it is necessary to restrict a model to only include the features relevant for explaining a variable of interest and discard those which are irrelevant. The difficulty lies in the set of relevant predictors being unknown to the modeller. The Lasso (or Least Absolute Shrinkage and Selection Operator) due to Tibshirani (1996) is the go-to estimator in this setting. It simultaneously selects a subset of relevant features and estimates the model by solving a convex optimization problem. The optimization problem involves the usual least squares objective function plus a term that penalizes the ℓ_1-norm of the coefficient vector. This penalty results in some of the coefficients being set exactly to zero, which functions as automatic model selection. If the goal is prediction or classification, such as determining whether an email is spam based on the words it contains, then the parameter affecting the strength

Email address: robert.garrard@adelaide.edu.au

September 7, 2017

of the penalty is set to minimize mean square error in an independent test data set.

The econometrician faces a similar problem but with the ultimate goal being causal inference on a treatment effect. Careful selection of which features enter an econometric model is important since omitting a relevant variable leads to inconsistent parameter estimates and including irrelevant variables leads to inflated standard errors, which weaken inference. Approaches to causal inference which require the researcher to have oracle-like knowledge with regard to which features are relevant has been heavily criticized, most notably by Leamer (1983) and Ioannidis (2005). The recent replication crisis in psychology (Open Science Collaboration, 2015), which has triggered similar attempts which fail to replicate experimental findings in economics (Camerer et al., 2016; Chang and Li, 2015), has raised questions as to whether there is a widespread use of 'p-hacking', a process in which a researcher selects variables to enter the model in order to engineer statistical significance where there is none, (Head et al., 2015). This comes despite a concerted effort to "take the con out of econometrics" (Leamer, 1983) in which applied econometrics experienced the so called credibility revolution (Angrist and Pischke, 2010) where focus turned to the plausible justification of correctly identifying a causal effect aided by instrumental variable (IV) and panel data methods. The Lasso offers a way in which econometric inference may be conducted without suspicion that a model has been cherry-picked by providing an automated and transparent procedure for model selection.

In this paper we present a selective review of studies detailing how the Lasso may be adapted to provide valid econometric inference in which we draw on both the statistical learning and econometrics literatures. The objective of this paper is to introduce the applied economist to Lasso methods as they relate to causal econometric inference by examining a few key papers addressing concepts relevant to current econometric practice.

In section 2 we define the Lasso estimator and address conditions for its consistency. Since the Lasso serves two functions, model selection and parameter estimation, we consider two disjoint notions of consistency. Zhao and Yu (2006) provide necessary and sufficient conditions for the Lasso to learn the true model; that is, which regressors have non-zero coefficients and which are irrelevant. Meinshausen and Yu (2009) characterize the conditions for the standard notion of consistency of an estimator, that the ℓ_2-norm difference between the estimator and the true parameter converge in probability to zero. We also consider a third kind of consistency due to Greenshtein and Ritov (2004), prediction consistency, which requires that the fitted values produced by the Lasso estimator converge in probabliity to the fitted values produced by the true parameter. We illustrate a simple proof of this kind of consistency to build intuition behind a variant of the Lasso due to Belloni et al. (2011), the square-root Lasso. The standard Lasso requires that we choose a tuning parameter proportional to the unknown standard deviation of the random disturbances. This can be estimated by running least squares or a pilot Lasso, but this is rather difficult in high dimensions. The square-root Lasso, with a simple modification to the Lasso objective function, achieves pivotal recovery of the parameter vector without requiring knowledge of the unknown standard deviation.

Section 3 discusses the construction of confidence intervals in the classical linear regression setting. Typically confidence intervals are constructed in one of two ways, either through characterizing the limiting distribution of an estimator and extracting the relevant quantiles, or by using the bootstrap (Efron, 1979). In the case where the dimension p is fixed, Knight and Fu (2000) characterize the limiting distribution of the Lasso estimator in terms of the argmin of a random function which does not admit a simple closed form. For a non-trivial choice of the tuning parameter the Lasso estimator is $\sqrt{n}$-consistent but experiences asymptotic bias for the non-zero elements of the true parameter vector. Furthermore, Knight and Fu (2000) conjecture that a naive bootstrap will fail to be consistent. Chatterjee and Lahiri (2010) prove that a residual bootstrap is in fact inconsistent but provide a modified bootstrap in Chatterjee and Lahiri (2011) which is not only consistent but may be used to consistently estimate the asymptotic bias and construct valid confidence intervals. For the high dimensional setting, where $p > n$, we discuss sample splitting techniques based on Wasserman and Roeder (2009).

Section 4 presents results on the use of the Lasso in instrumental variable (IV) regression. The standard two stage least squares (2SLS) estimator attempts to form an optimal combination of instruments in the first stage of the regression which predict the endogenous regressor. Those fitted values are then substituted for the endogenous regressor in the second stage after which OLS is performed. The many instruments problem refers to the use of too many instruments in the first stage generating unboudedly large variance in the second stage IV estimator. For this reason it is typically not feasible to introduce transformations of instruments in the first stage to attempt to capture non-linearities in the regression function. The Lasso benefits from producing a sparse predictor in which many of the estimated coefficients will be exactly zero, effectively removing irrelevent instruments from the model. Belloni et al. (2012) provide conditions under which Lasso in the first stage yields a consistent and asymptotically efficient IV estimator.

Section 5 concludes with a discussion.

2. The Lasso

2.1. Setup

Consider the following general regression model.

$$\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix, $\boldsymbol{y}$ is an $n \times 1$ response variable, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of mean zero sampling error, and $f(\cdot)$ is the conditional expectation function $f = \mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}]$. Given an i.i.d sample $(y_i, \boldsymbol{x}_i')_{i=1,\ldots,n}$, the goal of regression is to estimate the conditional expectation function (CEF), f. A typical approach is to assume that the CEF has a linear functional form.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

In the low dimensional setting, when $p \ll n$, the standard estimator is Ordinary Least Squares (OLS). The properties of this estimator and inferences based on it are well known when a particular set of assumptions are satisfied, most importantly that $\mathbb{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}] = 0$. In particular, the Gauss-Markov theorem declares the OLS estimator BLUE, the best linear unbiased estimator. That is, no linear unbiased estimator may have smaller variance. However, this comes with a caveat. No LUE applied *to the pair* $(\boldsymbol{y}, \boldsymbol{X})$ may have smaller variance. But if some of the regressors are irrelevant, that is they may be removed to form a design matrix $\tilde{\boldsymbol{X}}$ while maintaining the property that $\mathbb{E}[\tilde{\boldsymbol{X}}'\boldsymbol{\varepsilon}] = 0$, then the OLS estimator applied to $(\boldsymbol{y}, \tilde{\boldsymbol{X}})$ will have lower variance. While the estimator is still consistent with the inclusion of irrelevant regressors, ideally we would select only the regressors with non-zero coefficients, although this is usually not known a priori to the researcher.

Alternatively, we may find ourselves in the high dimensional setting where $p \gg n$. In this event the minimizer of the sum of square residuals is

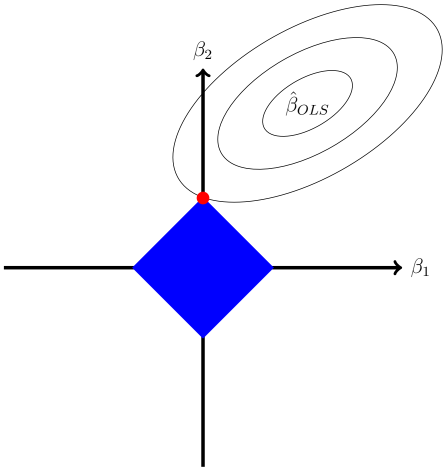


Figure 1: Geometry of the Lasso. Ellipses represent contours of the least squares objective function. The diamond represents the ℓ_1 constraint. Note that the constraint set has sharp corners, allowing estimates to be set exactly to zero.

not unique and the OLS estimator cannot be formed in the usual way, since the Gram matrix, $\frac{1}{n}\boldsymbol{X'X}$, is rank deficient. The event that $p \gg n$ might occur either due to the data set being of high dimensional nature or through the researcher not wishing to assume a linear functional form for the CEF. To capture non-linearities in the CEF the researcher may extend the design matrix to include a set of polynomial transformations and interactions of regressors. This can increase the dimension of the problem exponentially. In either setting, the Lasso offers a viable option.

2.2. Lasso

The Lasso, first proposed by Tibshirani (1996) and inspired by the non-negative garrote of Breiman (1995), solves the following convex optimization problem.

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta}}{\text{argmin}} \quad ||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{3}$$

where λ is a tuning parameter and $|| \cdot ||_q$ denotes the ℓ_q-norm. For penalized optimization problems it is common to standardize the data so that the solution is invariant to the choice of units. That is, the data are recentered and rescaled such that $\frac{1}{n}\sum_i y_i = 0$, $\frac{1}{n}\sum_i x_{i,j} = 0$, $\frac{1}{n}\sum_i x_{i,j}^2 = 1$. The problem above may be thought of as the Lagrangian form of the

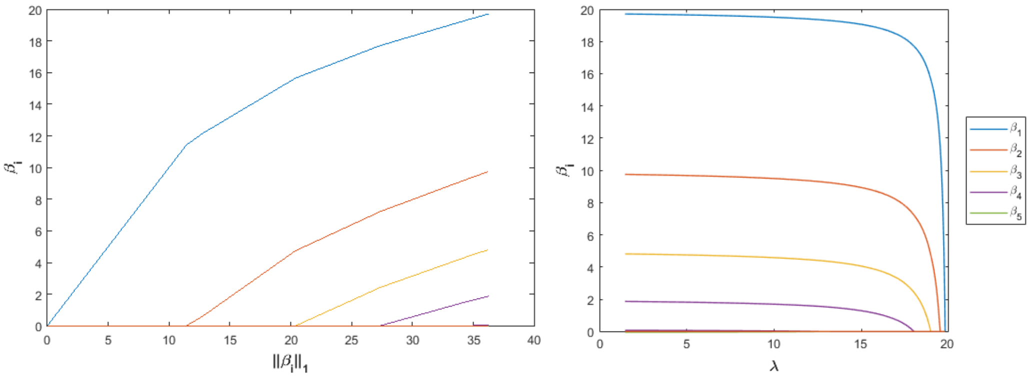


Figure 2: Lasso solution path.

optimization problem that minimizes the sum of square residuals subject to the constraint that the ℓ_1 norm of the solution is less than some amount. This ℓ_1 penalty serves to shrink the OLS estimates toward zero, and since the ℓ_1 ball has sharp corners, some of the entries in the solution may be set exactly to zero. Thus the Lasso simultaneously performs model selection and estimation. See figure 1 for the geometry of the problem.

The Lasso solution depends on the choice of the tuning parameter λ. If $\lambda = 0$, the constraint does not bind and the OLS solution is restored. If λ is sufficiently large, the penalty term dominates and all coefficients are set to 0. Choice of the value of the tuning parameter typically requires computing the whole solution path; that is, the solution to the lasso problem for many values of λ. This may be done efficiently with the LARS algorithm of Efron et al. (2004). Define the **active set**, $\mathcal{A}$, to be the set of coefficients that the Lasso estimates to be non-zero

$$\mathcal{A} = \{i \mid \hat{\boldsymbol{\beta}}_i \neq 0\} \tag{4}$$

and let $s = |\mathcal{A}|$ be the number of regressors in the active set.[1] Figure 2 illustrates the Lasso solution path for a model whose regressors are i.i.d $N(0, \frac{1}{2})$ random variables. In this illustration we have 50 observations and 100 regressors. The true parameter vector is $\boldsymbol{\beta} = [20, 10, 5, 2, 0, 0, 0, \dots]'$.

[1] Note that since both $\hat{\boldsymbol{\beta}}$ and $\mathcal{A}$ explicitly depend on λ they should be indexed as such. We abuse notation here for aesthetic reasons.

The Lasso path plots the estimated value of each parameter against the overall ℓ_1-norm of the solution vector. The leftmost part corresponds to a large λ such that all estimates are set to zero and the ℓ_1 norm of $\hat{\boldsymbol{\beta}}_{Lasso}$ is also zero. The rightmost part corresponds to $\lambda = 0$ and $\hat{\boldsymbol{\beta}}_{Lasso}$ returning a least squares solution. As we move from left to right, such that λ becomes less restrictive on the ℓ_1-norm of the estimator, we see that $\hat{\boldsymbol{\beta}}_1$, corresponding to the variable most correlated with $\boldsymbol{y}$, is the first to enter the active set, followed by the second most correlated variable, and so on.

When the objective is to minimize predictive risk, λ is usually chosen through cross-validation (Arlot and Celisse, 2010). This involves splitting the data set into k chunks. One chunk is held out, the model trained on the remaining $k-1$ chunks pooled together, and the hold out chunk is then used to estimate mean square prediction error. This is done for each of the k chunks being held out one at a time. The tuning parameter is chosen to minimize average prediction error. While this method for determining the tuning parameter is relatively simple and yields models with high predictive accuracy, it is unclear whether or not this method results in consistent estimation.

2.3. Consistency of the Lasso

Before discussing the conditions under which the Lasso is consistent, we must first precisely specify what notion of consistency to which we are referring. Note that the Lasso serves two functions: to choose a relevent subset of regressors, and to estimate coefficients for those regressors. This leads to two distinct notions of consistency. The former is typically called model selection consistency, and for the latter we use the usual notion of ℓ_2-consistency. Note that neither form of consistency implies the other.

Suppose that the true CEF is linear, as in equation (2), and denote $\mathcal{A}_0 = \{i \mid \boldsymbol{\beta}_i \neq 0\}$, with $s_0 = |\mathcal{A}_0|$.

Definition 1. An estimator $\hat{\boldsymbol{\beta}}$ is model selection consistent if

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{A} = \mathcal{A}_0\right] = 1 \tag{5}$$

This notion of consistency requires that the estimator consistently selects the exact set of regressors whose true coefficients are non-zero. The following condition for model selection consistency is due to Zhao and Yu (2006).

Let $\boldsymbol{X}$ be an $n \times p$ design matrix and $\boldsymbol{\beta}$ be the unknown vector of coefficients such that p and $\boldsymbol{\beta}$ are fixed as $n \to \infty$. Let $\boldsymbol{C}^n = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ be the Gram matrix of $\boldsymbol{X}$. Suppose, without loss of generality, that the first s_0 columns of $\boldsymbol{X}$ correspond to the true active variables and the remaining $p - s_0$ columns correspond to the irrelevant regressors. Then we may write the Gram matrix in block form.

$$\boldsymbol{C}^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix} \tag{6}$$

Similarly, define $\boldsymbol{\beta}_{(1)}$ to be the first s_0 coefficients, which are non-zero. Assuming C_{11} is invertible we may define the following condition.

Definition 2 (Irrepresentable Condition)**.**

$$|C_{21}^n (C_{11}^n)^{-1} \text{sign}(\boldsymbol{\beta}_{(1)})| < \mathbf{1}$$

Roughly speaking, the irrepresentable condition requires that variation in the irrelevant regressors cannot be accurately represented by a linear combination of relevant regressors. In other words, the irrelevant regressors are not “too correlated” with the relevant ones. This condition happens to be necessary and sufficient for allowing the Lasso to correctly identify the active set.

Theorem 1 (Zhao and Yu, 2006)**.** *Let p and $\boldsymbol{\beta}$ be fixed as $n \to \infty$, let $\boldsymbol{C}^n \to \boldsymbol{C}$ where $\boldsymbol{C}$ is positive definite, and $\frac{1}{n} \max_{1 \leq i \leq n} ((\boldsymbol{x}_i^n)'\boldsymbol{x}_i^n) \to 0$. Further, let the tuning parameter be selected such that $\lambda_n / n \to 0$ and $\lambda_n / n^{\frac{1+c}{2}} \to \infty$ for some $0 \leq c < 1$. Then the Lasso is model selection consistent if and only if there exists an N such that the irrepresentable condition holds for all $n > N$.*

Zhao and Yu (2006) also provide necessary and sufficient conditions for model selection consistency in the case where p and s_0 are allowed to grow with n and also illustrate a few simple cases in which the irrepresentable condition is guaranteed to hold. In general, the irrepresentable condition is a very stringent requirement that is unlikely to hold. As such, it is unreasonable to expect that the Lasso will recover the exact true active set asymptotically. However, the Lasso can learn a *superset* of the true active set, as we willl soon see. But first let us consider the second, and more familiar, form of consistency: ℓ_2-consistency of the coefficient vector.

Definition 3. An estimator $\hat{\boldsymbol{\beta}}$ is ℓ_2-consistent if

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2 \xrightarrow{p} 0 \quad \text{as } n \to \infty$$

Note that this version of consistency does not require that the active set contains no irrelevant regressors asymptotically, but it requires that the estimated coefficients for these regressors go to zero. Meinshausen and Yu (2009) provide the following conditions for ℓ_2-consistency.

Assume we have a linear model, as in equation (2), with $n \times p$ design matrix $\boldsymbol{X}$, unknown coefficient vector $\boldsymbol{\beta}$, and errors satisfying $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$, though the normality assumption may be relaxed to errors having sub-Gaussian tails. In this setup we will allow the dimension and the number of active regressors to grow with n. As such we will index the dimension p_n and the sparsity $s_n = |\{i \mid \boldsymbol{\beta}_i^n \neq 0\}|$.

The key assumption driving the result involves the notion of sparse eigenvalues, introduced by Donoho (2006). The m-sparse minimal eigenvalue of a matrix is the minimal eigenvalue of any $m \times m$ submatrix, and analogously for the m-sparse maximal eigenvalue.

Definition 4. Let $C = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. The m-sparse minimal and maximal eigenvalue of C are defined as

$$\phi_{min}(m) = \min_{v \,:\, ||v||_0 \leq \lceil m \rceil} \frac{v'Cv}{v'v} \qquad \phi_{max}(m) = \max_{v \,:\, ||v||_0 \leq \lceil m \rceil} \frac{v'Cv}{v'v}$$

We may now state a sufficient condition for ℓ_2-consistency.

Theorem 2 (Meinshausen and Yu, 2009)**.** *Assume that there exist constants* $0 < \kappa_{min} \leq \kappa_{max} < \infty$ *such that*

$$\liminf_{n \to \infty} \phi_{min}(s_n \log n) \geq \kappa_{min}$$

$$\limsup_{n \to \infty} \phi_{max}(s_n + \min\{n, p_n\}) \leq \kappa_{max}$$

Then for $\lambda \propto \sigma\sqrt{n \log p_n}$*,* $\exists M > 0$ *such that, with probability converging to* 1 *for* $n \to \infty$

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2 \leq M\sigma\sqrt{\frac{s_n \log p_n}{n}}$$

Therefore ℓ_2-consistency obtains if $\frac{s_n \log p_n}{n} \to 0$. This is satisfied in the cases where $\boldsymbol{\beta}$ and p are fixed, or when $\boldsymbol{\beta}$ is fixed and the dimension grows at a rate such that $\log p_n = o(n)$. ℓ_2-consistency of the Lasso implies the following screening property.

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{A}_0 \subseteq \mathcal{A}\right] = 1 \tag{7}$$

While the Lasso may not precisely recover the true active set, with high probability it will include all relevant regressors and a few irrelevent regressors, whose estimated coefficents will be tending to zero.

To summarize, if we choose the tuning parameter such that $\lambda \propto \sigma\sqrt{n \log p_n}$; the true parameter vector $\boldsymbol{\beta}$ is sufficiently sparse, meaning that not too many coefficients are non-zero; and the regressors are not "too correlated", resulting in the above restricted eigenvalue conditions holding; then the Lasso estimates a model containing all relevant regressors with high probability and consistently estimates the coefficient vector $\boldsymbol{\beta}$. The issue here is that selection of the tuning parameter requires knowledge of the unkown standard deviation σ. In the low-dimensional setting, this can easily be estimated, but when $p \gg n$ this can be especially difficult.

2.4. Prediction Consistency

Before continuing on to a variant of the Lasso which does not require estimating the unknown variance in order to choose the tuning parameter, it will be informative to first consider an alternative notion of consistency and an accompanying proof which is rather straightforward. This proof will help to build intuition for the next section.

Predictive consistency, also called persistency by Greenshtein and Ritov (2004), is similar to ℓ_2-consistency for the coefficient vector but relates to the 2-norm of the fitted value vector.

Definition 5. An estimator $\hat{\boldsymbol{\beta}}$ is prediction consistent if

$$\frac{1}{n}\mathbb{E}||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \xrightarrow{p} 0$$

This notion of consistency is much more mild than those in section 2.3. It does not require that the Lasso learn the true coefficient vector $\boldsymbol{\beta}$, it only requires that the fitted values converge to the truth. In what follows, let us assume that disturbances are Gaussian and homoscedastic, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$.

Supposing we possessed an oracle which knew the true active set, $\mathcal{A}_0$, we could perform OLS to achieve maximum efficiency. The oracle estimator has the following rate of convergence in the prediction norm.

$$\frac{1}{n}\mathbb{E}||\boldsymbol{X}\hat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 = \sigma^2\frac{s_0}{n} \tag{8}$$

To establish prediction consistency for the Lasso, we begin with the so-called Basic Inequality.

$$||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2 + \lambda||\hat{\boldsymbol{\beta}}||_1 \leq ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{9}$$

This holds since by definition $\hat{\boldsymbol{\beta}}$ is the argmin of the Lasso objective function. Rearranging

$$||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq \lambda\left(||\boldsymbol{\beta}||_1 - ||\hat{\boldsymbol{\beta}}||_1\right) \tag{10}$$

Let $Q(\boldsymbol{b}) = ||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||_2^2$ be the least squares part of the objective function and define the score function to be the gradient of this function evaluated at the true parameter $\boldsymbol{\beta}$.

$$\boldsymbol{S} = \nabla Q(\boldsymbol{\beta}) = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = -2\boldsymbol{X}'\boldsymbol{\varepsilon} \tag{11}$$

Expanding the left hand side of equation (10) and rearranging gives

$$||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq -\boldsymbol{S}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \lambda\left(||\boldsymbol{\beta}||_1 - ||\hat{\boldsymbol{\beta}}||_1\right) \tag{12}$$

Using Holder's inequality we have

$$||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq ||\boldsymbol{S}||_\infty||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_1 + \lambda\left(||\boldsymbol{\beta}||_1 - ||\hat{\boldsymbol{\beta}}||_1\right) \tag{13}$$

Therefore if we choose λ such that $\lambda > ||\boldsymbol{S}||_\infty$ with high probability, then, applying the triangle inequality

$$||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq 2\lambda||\boldsymbol{\beta}||_1 \tag{14}$$

To pick $\lambda > 2||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty$ we may note that for $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ we may bound the score by

$$||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty \leq 1.01\sigma\sqrt{n}\sqrt{2\log p} \tag{15}$$

with probability tending to one. Thus we obtain

$$\frac{1}{n}||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq 4.04\sqrt{2}\sigma\sqrt{\frac{\log p}{n}}||\boldsymbol{\beta}||_1 \tag{16}$$

for $\lambda \propto \sigma\sqrt{n\log p}$. Thus we obtain prediction consistency as long as the ℓ_1-norm of the coefficient vector and the dimension do not grow too quickly.

Note that the rate of convergence for the oracle estimator, equation (8), is $O(n^{-1})$, whereas the rate of convergence for the Lasso is $O(n^{-1/2})$. This is the so-called slow rate for the Lasso, which requires very few assumptions on the design matrix to prove. If we assume more structure on the design, similar to the restricted eigenvalue condition in section 2.3, we can reacquire the fast rate of $O(n^{-1})$.

The key concept to take away from this derivation is that the difference in the least squares parts of the basic inequality may be written in terms of a score function. Appropriate selection of the tuning parameter requires that the tuning parameter be sufficiently large so as to dominate the sup-norm of the score function, but not so large that the inequality fails to be tight. In this case we exploited the normality assumption to bound the sup-norm of the score function, which is a function of σ, with high probability. This resulting in a tuning parameter that must be selected such that $\lambda \propto \sigma\sqrt{n\log p}$. The difficulty here, as for the consistency result in section 2.3, is that σ is unknown and must be estimated.

2.5. The Square-Root Lasso

The square-root Lasso, or $\sqrt{\text{Lasso}}$, due to Belloni et al. (2011) is a modification to the Lasso which allows for pivotal recovery of the parameter vector. That is, it does not require knowledge or estimation of the unknown standard deviation in order to choose the tuning parameter. The objective function for the $\sqrt{\text{Lasso}}$ takes the form

$$\hat{\boldsymbol{\beta}}_{\sqrt{\text{Lasso}}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \quad \sqrt{\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2} + \frac{\lambda}{n}||\boldsymbol{\beta}||_1 \tag{17}$$

with the tuning parameter chosen such that $\lambda = c\sqrt{n}\Phi^{-1}(1-\alpha/2p)$, where $c > 1$ is a constant, α is the desired significance level, and Φ denotes the standard normal CDF. The key difference here is that the square-root of the least squares part of the objective function is taken. While this may appear to be a minor modification, the resulting estimator not only achieves ℓ_2-consistency for the parameter vector under mild conditions, but under slightly stronger conditions may be generalized to an arbitrary distribution of noise.[2] The conditions for consistency are similar to the restricted eigenvalue conditions of Meinshausen and Yu (2009) and give the following inequality

$$||\hat{\boldsymbol{\beta}}_{\sqrt{\text{Lasso}}} - \boldsymbol{\beta}||_2 \leq M\sigma\sqrt{\frac{s_n \log(2p/\alpha)}{n}} \tag{18}$$

with probability approaching $1-\alpha$.

For intutition, let us rewrite the model to be estimated as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sigma\boldsymbol{u} \tag{19}$$

where $\boldsymbol{u} \sim N(0, I)$. Let $Q(\boldsymbol{b}) = ||\boldsymbol{y} - \boldsymbol{Xb}||_2^2$ be the usual least squares part of the objective function. Analogous to section 2.4, we must choose λ so as to bound the score of $Q^{1/2}$. Using the chain rule

$$\nabla Q^{1/2}(\boldsymbol{b}) = \frac{\frac{1}{2}\nabla Q(\boldsymbol{b})}{Q^{-1/2}(\boldsymbol{b})} = \frac{\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{Xb})}{||\boldsymbol{y} - \boldsymbol{Xb}||_2} \tag{20}$$

Evaluating at $\boldsymbol{\beta}$ and substituting in the DGP gives

$$\nabla Q^{1/2}(\boldsymbol{\beta}) = \frac{-\sigma\boldsymbol{X}'\boldsymbol{u}}{||\sigma\boldsymbol{u}||_2} = \frac{-\boldsymbol{X}'\boldsymbol{u}}{||\boldsymbol{u}||_2} \tag{21}$$

which is pivotal with respect to σ. Thus the λ chosen to dominate the score is not a function of the unknown variance.

There are many modifications to the standard Lasso estimator that may be of interest to the econometrician which we do not elaborate on here. The adaptive Lasso (Zou, 2006), which achieves asymptotic unbiasedness by weighting each coefficient in the penalty term by its least squares estimate;

[2] Recall that Meinshausen and Yu (2009) required Gaussian or sub-Gaussian noise.

the elastic net (Zou and Hastie, 2005), which uses a penalty that is a linear combination of the ℓ_1-norm and ℓ_2-norm of the coefficient vector achieving better predictive properties when regressors are highly correlated; Fan and Li (2001), which introduce the SCAD penalty (smoothly clipped absolute deviation); the group Lasso (Yuan and Lin, 2006), which allows for categorical variables; the fused Lasso (Tibshirani et al., 2005), which allows for regressors to be ordered; and the post-Lasso (Belloni and Chernozhukov, 2009), which discards the Lasso coefficient estimates and performs least squares on the selected model.

3. Classical Inference

While an estimator may have desirable properties, such as consistency, unbiasedness, or efficiency, by its nature it is a random variable which is subject to sampling error. Were a different sample drawn, a different value for the estimator would have been realized by the fact that we sample a randomly selected subset of the population. The objective of statistical inference is to quantify the uncertainty due to sampling error. Typically this involves either conducting a hypothesis test to determine whether or not an effect size is significantly different from zero, or constructing a confidence set to determine a sub-region of the parameter space where the true parameter is likely to reside. Since confidence sets and hypothesis tests are tightly linked, where one may be inverted to construct the other, in the rest of this section we will only consider the construction of confidence sets.

A $1-\alpha$ confidence set is a set $\mathcal{C} \subset \mathbb{R}^p$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\boldsymbol{\beta} \in \mathcal{C}\right) \geq 1-\alpha \tag{22}$$

That is, it is a procedure for generating a set, $\mathcal{C}$, which traps the true parameter, $\boldsymbol{\beta}$, with some specified probability. For illustration, consider the following example of a 90% confidence set. Let Y be a Bernoulli random variable with $\mathbb{P}(Y=1)=0.9$ and $\mathbb{P}(Y=0)=0.1$. Construct the set $\mathcal{C}$ such that

$$\mathcal{C} = \begin{cases} \emptyset & \text{if } Y=0 \\ \mathbb{R}^p & \text{if } Y=1 \end{cases} \tag{23}$$

This is a valid confidence set since it traps the true parameter with $\mathbb{P}(\boldsymbol{\beta} \in \mathcal{C}) = 0.9$. This trivial example shows that having coverage as advertized is not the only desirable property of a confidence set. Typically we want a confidence set which contains, or even is centered at, $\hat{\boldsymbol{\beta}}$, a point estimator for the true parameter. This is typically accomplished in one of two ways. The first requires us to characterize the asymptotic distribution of the estimator.

Consider the standard linear model in equation (1) with $p < n$ so that the standard OLS estimator may be applied.

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} \tag{24}$$

Substituting the DGP and rearranging gives

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \boldsymbol{x}_i'\boldsymbol{\varepsilon}_i \tag{25}$$

Assuming $\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i' \to \boldsymbol{C}$, where $\boldsymbol{C}$ is invertible, and further supposing $\boldsymbol{\varepsilon}_i$ are homoscedastic, by the CLT we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \tag{26}$$

Thus to construct a confidence set for $\boldsymbol{\beta}_i$ we may use the following set

$$\mathcal{C} = \left(\hat{\boldsymbol{\beta}}_i - z_{\alpha/2}\frac{\hat{\sigma}(\boldsymbol{X}'\boldsymbol{X})_{ii}^{-1/2}}{\sqrt{n}},\ \hat{\boldsymbol{\beta}}_i + z_{\alpha/2}\frac{\hat{\sigma}(\boldsymbol{X}'\boldsymbol{X})_{ii}^{-1/2}}{\sqrt{n}}\right) \tag{27}$$

Where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution and $\hat{\sigma}$ is a consistent estimator of σ. Since we know that the limiting distribution of $\hat{\boldsymbol{\beta}}_i$ is Gaussian, the above interval covers $\boldsymbol{\beta}_i$ with probability $1-\alpha$ asymptotically.

The main alternative to characterize the limiting distribution of an estimator is to use the bootstrap, due to Efron (1979), which attempts to simulate the sampling distribution of $\hat{\boldsymbol{\beta}}$. Let

$$H_n(\tau) = \mathbb{P}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \tau\right) \tag{28}$$

be the unknown sampling distribution for $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. The objective of the bootstrap is to approximate this distribution with some function, $\hat{H}_n(\tau)$.

Suppose that the data are distributed according to the law $(y_i, \boldsymbol{x}_i') \overset{i.i.d}{\sim} F$. Let $\hat{F}_n$ be the empirical distribution of a sample of size n. The well known Glivenko-Cantelli theorem states that

$$\sup_x |\hat{F}_n(x) - F(x)| \overset{a.s}{\to} 0 \tag{29}$$

That is, the empirical distribution of the sample converges uniformly to the distribution of the true unknown data generating process. If we were able to repeatedly draw from the unknown DGP, we could easily construct $H_n(\tau)$ simply by taking B draws from F, constructing $\hat{\boldsymbol{\beta}}$ for each, and looking at the distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ as $B \to \infty$. This is infeasible since F is unknown. The ingenuity of the bootstrap is to use $\hat{F}_n$ as a proxy for F.

In what follows we will consider the same setup as above where $\hat{\boldsymbol{\beta}}$ is the OLS estimator applied to the sample $(y_i, \boldsymbol{x}_i')$, $i = 1, \ldots, n$. Here we will consider the residual bootstrap, which is standard for the case when the design matrix is non-random.

Algorithm 1 Residual Bootstrap

```
Let $e_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, $\bar{e} = n^{-1}\sum e_i$, and let $\{e_i - \bar{e}\}$ be the set of centered residuals
for $j = 1$ to $B$ do
    Draw with replacement a sample of size $n$ from the centered residuals, $\{e_i^*, i = 1, \ldots, n\}$
    Form the bootstrap data set $y_i^* = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + e_i^*$
    Construct the bootstrap OLS estimator $\hat{\boldsymbol{\beta}}_j^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}^*$
end for
Let $\hat{H}_n(\tau) = \mathbb{P}_*\left(\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \leq \tau\right)$
```

From Freedman (1981) we have

Theorem 3. *Let $\boldsymbol{X}$ be a non-random design and e_i be i.i.d with $\mathbb{E}[e_i^2] = \sigma^2$ and $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \to \boldsymbol{V}$, which is positive definite. Then the residual bootstrap attains*

$$\sup_\tau |\hat{H}_n(\tau) - H(\tau)| \overset{p}{\to} 0$$

Thus if we set $t_{\alpha/2} = \hat{H}_n^{-1}(\alpha/2)$ and $t_{1-\alpha/2} = \hat{H}_n^{-1}(1-\alpha/2)$, we can construct the following approximate $1-\alpha$ confidence interval for $\boldsymbol{\beta}$.

$$\mathcal{C} = \left(\hat{\boldsymbol{\beta}} - \frac{t_{1-\alpha/2}}{\sqrt{n}}, \hat{\boldsymbol{\beta}} - \frac{t_{\alpha/2}}{\sqrt{n}}\right) \tag{30}$$

3.1. Limiting Distribution of the Lasso Estimator

Knight and Fu (2000) consider the limiting distribution of and a bootstrap for the Lasso in the low dimensional setting, ie $p < n$, with p fixed. Consider a linear model, as in equation (2), where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d variables with mean 0 and constant variance σ^2. Let $\hat{\boldsymbol{\beta}}_n$ be the Lasso estimator and λ_n be the Lasso penalty. Assume the following standard conditions on the design matrix

$$\boldsymbol{C}_n = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \to \boldsymbol{C} \qquad \frac{1}{n}\max_{1\leq i\leq n} \boldsymbol{x}_i'\boldsymbol{x}_i \to 0 \tag{31}$$

where $\boldsymbol{C}$ is non-negative definite.

Theorem 4. *If $\boldsymbol{C}$ is non-singular and $\lambda_n/n \to \lambda_0 \geq 0$, then $\hat{\boldsymbol{\beta}}_n \xrightarrow{p}$ argmin(Z) where*

$$Z(\boldsymbol{\phi}) = (\boldsymbol{\phi} - \boldsymbol{\beta})'\boldsymbol{C}(\boldsymbol{\phi} - \boldsymbol{\beta}) + \lambda_0 \sum_{j=1}^{p} |\phi_j|$$

Thus if $\lambda_n = o(n)$, argmin$(Z) = \boldsymbol{\beta}$, and so $\hat{\boldsymbol{\beta}}_n$ is consistent.

Therefore for consistency of the Lasso in the *low dimensional* setting, it is sufficient that $\lambda_n/n \to 0$. Constrast this with the high dimensional setting of Meinshausen and Yu (2009) discussed in section 2.3. In that setting it was required that $\lambda \propto \sigma\sqrt{n \log p}$. Thus for fixed p, $\lambda/n \to 0$, and the requirement that $\boldsymbol{C}$ be non-singular satisfies the restricted eigenvalue conditions. Thus theorem 4 is a special case of Meinshausen and Yu (2009).

While we require $\lambda = o(n)$ for consistency, in order for the Lasso estimator to converge to a non-trivial asymptotic distribution the tuning parameter must grow more slowly than $\lambda_n = O(n)$.

Theorem 5. *If $\boldsymbol{C}$ is non-singular and $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \textit{argmin}(V)$$

where

$$V(\boldsymbol{u}) = -2\boldsymbol{u}'\boldsymbol{W} + \boldsymbol{u}'\boldsymbol{C}\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} [u_j sgn(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)]$$

where $\boldsymbol{W}$ *has a* $N(\mathbf{0}, \sigma^2 \boldsymbol{C})$ *distribution.*

Suppose $\lambda_0 = 0$. Then minimizing $V(\boldsymbol{u})$ with respect to $\boldsymbol{u}$ yields

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \boldsymbol{C}^{-1}) \tag{32}$$

which is the asymptotic distribution of the OLS estimator. This is undesirable since $\lambda_0 = 0$ means that the Lasso would not perform any model selection asymptotically. So we must have $\lambda_0 > 0$ for the Lasso to function as a selection operator. The tradeoff is in that large λ_0 is required to correctly select the zeros elements of $\boldsymbol{\beta}$, but the bias in the non-zero elements of $\hat{\boldsymbol{\beta}}_n$ is proportional to λ_0. Despite $\hat{\boldsymbol{\beta}}_n$ being consistent for $\boldsymbol{\beta}$, when scaled by $\sqrt{n}$ the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is not centered at $\boldsymbol{\beta}$. This problem of bias not disappearing fast enough asymptotically is similar to that found in kernel density estimation. Thus confidence intervals centered at $\hat{\boldsymbol{\beta}}_n$ formed from the asymptotic distribution in theorem 5 will not attain correct coverage. Even supposing they were centered correctly, the asymptotic distribution of the Lasso is still difficult to extract critical values from since it is characterized as the argmin of a function. Closed form expressions for this distribution are not available except in special cases, such as for orthogonal design. This leaves characterizing the limiting distribution of the Lasso estimator as an infeasible method for constructing confidence intervals.

3.2. Bootstrapping the Lasso

The bootstrap would appear to be the prime candidate for contstructing intervals since it does not rely on knowledge of the asymptotic distribution of an estimator. However, Knight and Fu (2000) sketch out why the residul bootstrap will fail to be consistent for the Lasso estimator. Letting $H(\cdot)$ and $\hat{H}_n(\cdot)$ be the limiting distribution and bootstrap estimator defined previously, this property was later formalized by Chatterjee and Lahiri (2010) in the following theorem.[3]

[3] We have merged Theorem 3.1 and Corollary 3.2 from Chatterjee and Lahiri (2010) for brevity.

Theorem 6. *Suppose that*

(C.1) $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{x}_i \to \boldsymbol{C}$ *which is positive definite. Further*

$$n^{-1}\sum_{i=1}^{n} ||\boldsymbol{x}_i||_2^3 = O(1) \text{ as } n \to \infty$$

(C.2) $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$

(C.3) The errors $\{\varepsilon_i\}_{i=1}^n$ *are iid with* $\mathbb{E}[\varepsilon_i] = 0$ *and* $Var(\varepsilon_i) = \sigma^2$.

If $\{j \mid \boldsymbol{\beta}_j = 0\}$ *is non-empty and* $\lambda_0 > 0$, *then*

$$\varrho\left(\hat{H}_n(\cdot), H(\cdot)\right) \overset{p}{\not\to} 0 \text{ as } n \to \infty$$

Where ϱ *is the Prohorov metric.*

While it is unfortunate that the bootstrap fails to be consistent exactly when we need it due to the complexity of the limiting distribution, this should come at no surprise. From Bickel and Freedman (1981) and Andrews (1999, 2000) we know that the bootstrap will often fail to be consistent in the event that the distribution being bootstrapped is not continuous. The key reason to use the Lasso is for its model selection property. That is, the limiting distribution for the zero elements of $\boldsymbol{\beta}$ will have a point mass at zero, making the bootstrap distribution discontinuous. In particular, Chatterjee and Lahiri (2010) show that failure of the bootstrap in this instance is due to the Lasso failing to capture the sign of zero components with sufficient accuracy. While the Lasso attains the correct sign for non-zero components with high probability, it assigns both positive and negative signs to zero coefficients with positive probability.

Chatterjee and Lahiri (2011) propose a modified residual bootstrap able to consistently estimate the Lasso limiting distribution. The key modification is that coefficients are hard-thresholded in such a way that zero coefficients are given a sign of zero with high probability. Since the Lasso is $\sqrt{n}$-consistent, fluctuations of $\hat{\boldsymbol{\beta}}$ around the true value are of order $n^{-1/2}$. Thus we can achieve correct signs for zero elements with high probability by thresholding estimates within a neighborhood of order $n^{-1/2}$. Let $\{a_n\}$ be a sequence of real numbers such that

$$a_n + \left(n^{-1/2}\log n\right) a_n^{-1} \to 0 \text{ as } n \to \infty \tag{33}$$

For example, $a_n = cn^{-\delta}$ for $c \in (0, \infty)$ and $\delta \in (0, \frac{1}{2})$. Construct the modified Lasso estimator $\tilde{\boldsymbol{\beta}}$ by hard-thresholding at a_n.

$$\tilde{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_j I\left(|\hat{\boldsymbol{\beta}}_j| \geq a_n\right) \tag{34}$$

For non-zero components the estimated coefficient will be larger than a_n in absolute value with high probability, so the Lasso estimate and modified Lasso estimate will agree. However, for zero elements the estimated coefficient will be in the interval $[-a_n, a_n]$ with high probability and so will be thresholded to zero. Thus for large n, the signs of the components for the modified Lasso and the true parameter vector will agree. The modified bootstrap proceeds as follows.

Algorithm 2 Modified Residual Bootstrap

```
Let $r_i = y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}$, $\bar{r} = n^{-1}\sum r_i$, and let $\{r_i - \bar{r}\}$ be the set of centered residuals
for $j = 1$ to $B$ do
    Draw with replacement a sample of size $n$ from the centered residuals, $\{r_i^{**}, i = 1, \ldots, n\}$
    Form the bootstrap data set $y_i^{**} = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + r_i^{**}$
    Construct the bootstrap Lasso estimator $\hat{\boldsymbol{\beta}}_j^{**} = \underset{\boldsymbol{u}}{\text{argmin}}||\boldsymbol{y}^{**} - \boldsymbol{X}\boldsymbol{u}||_2^2 + \lambda||\boldsymbol{u}||_1$
    Let $\boldsymbol{T}_n^{**} = \sqrt{n}(\hat{\boldsymbol{\beta}}^{**} - \tilde{\boldsymbol{\beta}})$
end for
Let $\tilde{H}_n(\tau) = \mathbb{P}_*\left(\boldsymbol{T}_n^{**} \leq \tau\right)$
```

Which yields

Theorem 7. *Under the same conditions as theorem 6*

$$\varrho\left(\tilde{H}_n(\cdot), H(\cdot)\right) \to 0 \text{ as } n \to \infty \text{ with probability 1}$$

Furthermore

Theorem 8. *Under the same conditions as theorem 6*

$$\mathbb{E}_*[\sqrt{n}(\hat{\boldsymbol{\beta}}^{**} - \tilde{\boldsymbol{\beta}})] \to \mathbb{E}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$$

with probability 1.

Thus the asymptotic bias of the Lasso may be consistently bootstrapped. Letting $\hat{t}_n(1-\alpha)$ denote the $1-\alpha$ quantile of the bootstrap distribution of $||\boldsymbol{T}_n^{**}||_2$, we may set

$$\mathcal{C}_{n,1-\alpha} = \{\boldsymbol{t} \in \mathbb{R}^p \mid ||\boldsymbol{t} - \hat{\boldsymbol{\beta}}||_2 \leq n^{-1/2}\hat{t}_n(1-\alpha)\} \tag{35}$$

Corollary 1. *Suppose that* $\{j \mid \boldsymbol{\beta}_j = 0\}$ *is non-empty. Then for all* $\alpha \in (0,1)$

$$\mathbb{P}(\boldsymbol{\beta} \in \mathcal{C}_{n,1-\alpha}) \to 1-\alpha \textit{ as } n \to \infty$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$*.*

Thus the set in *equation* (35) functions as an approximate $1-\alpha$ confidence interval.

3.3. Sample Splitting

The two standard approaches to inference, characterizing the limiting distribution of an estimator and bootstrapping, both fail in the high dimensional setting. While there has been much headway into constructing confidence intervals and conducting hypothesis tests in high dimensions, such as Meinshausen et al. (2009), Bhlmann (2013), Zhang and Zhang (2014), Lockhart et al. (2014), Javanmard and Montanari (2014), Van de Geer et al. (2014) and Meinshausen (2015), in this section we will consider a method based on sample splitting implicit in Wasserman and Roeder (2009).[4] The basic idea is quite intuitive.

Take a sample of size n, $\mathcal{D} = (y_i, \boldsymbol{x}_i')_{i=1}^n$ which may be either high or low dimensional in nature. For simplicity, assume n is even. Split the sample randomly into two sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, each of size $n/2$, such that

$$\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset \qquad \mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$$

Use $\mathcal{D}_1$ to perform model selection and use $\mathcal{D}_2$ to conduct inference. For example, we might run the Lasso on $\mathcal{D}_1$ to estimate the active set $\mathcal{A}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{A}_0 \subseteq \mathcal{A}\right] = 1 \tag{36}$$

[4] See Dezeure et al. (2015) for a review of high dimensional inference.

Discard the coefficients from the Lasso and on $\mathcal{D}_2$ construct the OLS estimator for the screened model

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{X}_{\mathcal{A}}'\right)^{-1}\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{y} \tag{37}$$

from which we may proceed with inference as usual.

We do not encounter problems with post-selection inference since the data we use to perform inference was not used to select the model, so p-values and confidence intervals formed will attain their advertized size and coverage. While this concept is powerful in its simplicity and generality, it has yet to catch on in applied econometrics. This is likely due to the feeling that we are "throwing away" half our data, which is a big sacrifice to the efficiency of any estimator we might wish to use. However, the gain from doing so is that we can more directly capture non-linearities in the true CEF by including many transformations of regressors, which would render OLS rather inefficient, and we may automate the model selection procedure.

4. Instrumental Variables

In the classical linear regression setting, consistent estimation, and therefore valid inferences, relies crucially on the exogeneity assumption that disturbances are uncorrelated with the regressors: $\mathbb{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}] = 0$. While this assumption may be reasonable in physical and medical sciences, wherein a treatment is randomly assigned by the experimenter, in economic sciences this is something of a tall order. An economist is typically only able to observe a pair of outcomes and regressors, $(\boldsymbol{y}, \boldsymbol{X})$, randomly generated by some unknown data generating process. The inability to directly control the treatment variable in $\boldsymbol{X}$ makes the exogeneity assumption implausible in most cases.

The method of instrumental variables allows us to regain consistent estimation and valid inferences despite endogenous regressors. Let us assume that the true regression function is linear such that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{38}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix of endogenous regressors such that for each i, $\mathbb{E}[\boldsymbol{x}_i'\boldsymbol{\varepsilon}] \neq 0$. Without loss of generality we are assuming that all regressors in the main equation are endogenous, since equation (38) may be interpreted as the transformed variables after any exogenous variables have been partialed out with the appropriate projection. Due to endogeneity, the usual OLS estimator will fail to be consistent.

Suppose we had at our disposal an $n \times p$ matrix of instruments, $\boldsymbol{Z}$, with $p \geq k$ such that $\mathbb{E}[\boldsymbol{Z}'\boldsymbol{\varepsilon}] = 0$ but $\mathbb{E}[\boldsymbol{Z}'\boldsymbol{X}] \neq 0$. That is, these variables are correlated with the endogenous regressors but do not themselves belong in the main equation. These instruments are typically variables for which the exogeneity assumption is much more plausible, usually due to being randomly assigned by nature (such as your date of birth).

Suppose further that we have a function $G(\boldsymbol{Z}) : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times k}$ where $rank(G(\boldsymbol{Z})'\boldsymbol{X}) = k$.[5] The standard IV estimator is

$$\hat{\boldsymbol{\beta}}_{IV} = \left(G(\boldsymbol{Z})'\boldsymbol{X}\right)^{-1} G(\boldsymbol{Z})'\boldsymbol{y} \tag{39}$$

which has the following desirable property.

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}\right) \xrightarrow{d} N(0, \boldsymbol{\Sigma}^{-1}_{G(\boldsymbol{Z})\boldsymbol{X}} \boldsymbol{\Sigma}_{G(\boldsymbol{Z})G(\boldsymbol{Z})} \boldsymbol{\Sigma}'^{-1}_{G(\boldsymbol{Z})\boldsymbol{X}}) \tag{40}$$

Where $\mathbb{E}[G(\boldsymbol{Z})'\boldsymbol{X}] = \boldsymbol{\Sigma}_{G(\boldsymbol{Z})\boldsymbol{X}}$ and $\mathbb{E}[G(\boldsymbol{Z})'G(\boldsymbol{Z})] = \boldsymbol{\Sigma}_{G(\boldsymbol{Z})G(\boldsymbol{Z})}$. The optimal estimator in the minimum asymptotic variance sense is obtained by setting $G(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Z}]$, the CEF of $\boldsymbol{X}$ on $\boldsymbol{Z}$. If we further assume that the noise is homoscedastic, such that $\text{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{Z}) = \sigma^2 \boldsymbol{I}$, then we obtain

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}\right) \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}_{G(\boldsymbol{Z})G(\boldsymbol{Z})}) \tag{41}$$

which is the semiparametric efficiency bound of Newey (1990). However, the above efficient estimator is infeasible as the CEF, $G(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Z}]$, is unknown and must be estimated. As in the classical regression setting, the standard method involves assuming that G has a linear functional form.

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{\nu} \tag{42}$$

where $\boldsymbol{\Pi}$ is a $p \times k$ matrix of coefficients and $\mathbb{E}[\boldsymbol{\nu} \mid \boldsymbol{Z}] = 0$. This approach yields the two stage least squares (2SLS) estimator. The CEF is estimated by performing OLS in the 1st stage regression, equation (42).

[5] Note that when $p = k$ typically G is the identity function, $G(\boldsymbol{Z}) = \boldsymbol{Z}$.

The fitted values are used to approximate G, which is then used as the optimal instrument in the 2nd stage, equation (38) to yield

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{y} \tag{43}$$

where $\boldsymbol{P_Z} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ is the orthogonal projector into the column space of $\boldsymbol{Z}$. The 2SLS estimator has limiting distribution

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(0, \sigma^2\mathbb{E}[\boldsymbol{Z}'\boldsymbol{P_Z}\boldsymbol{Z}]^{-1}\right) \tag{44}$$

for which the asymptotic variance is consistently estimable.

The first stage of a 2SLS regression and the linear regression model in the classical setting are analogous: they each attempt to estimate a conditional expectation function by imposing a linear functional form and they each suffer from high variance if they include too many irrelevant variables. The tradeoff remains that the assumption of linearity is implausible, but including too many regressors to capture non-linearity (polynomial transformations and interactions) inflates the variance of the estimator to an unacceptable level. Several attempts have been made at capturing non-linearities without the accompanying variance inflation, such as Kloek and Mennes (1960), Amemiya (1966), Donald and Newey (2001), Chamberlain and Imbens (2004), Bai and Ng (2009), Caner (2009), and Okui (2011).

4.1. Lasso in the First Stage

Suppose that $\boldsymbol{Z}$, the $n \times p$ matrix of instruments, has large dimension p. This could either be because we are in the multiple-instrument setting, or we may wish to include polynomial transformations of the instruments to capture non-linearities in the CEF, $G(\boldsymbol{Z})$. If the representation of the CEF is sparse, that is only a few, but unknown set of elements of $\boldsymbol{\Pi}$ in equation (42) are non-zero, then the natural inclination would be to use the Lasso estimator in the first stage regression. In what follows we present the results from Belloni et al. (2012). These results are particularly novel as they hold for non-Gaussian heteroscedastic disturbances. We switch from matrix notation to indexing by i and n for convenience.

Suppose we have the following linear IV model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \qquad i = 1, \ldots, n \tag{45}$$

$$x_{i\ell} = \boldsymbol{z}_i'\boldsymbol{\delta}_\ell + \nu_{i\ell} \qquad \ell = 1, \ldots, k \quad i = 1, \ldots, n \tag{46}$$

where $\boldsymbol{x}_i'$ is a k-vector of endogenous variables, $\boldsymbol{z}_i'$ is a p-vector of instruments, $\boldsymbol{\delta}_l$ is an unknown vector of coefficients for constructing an instrument for the l-th endogenous variable, $\mathbb{E}[\varepsilon_i|\boldsymbol{z}_i] = 0$ and $\mathbb{E}[\nu_i|\boldsymbol{z}_i] = 0$ across all i and n. Note that the number of instruments may be much greater than the number of observations, $p > n$.

The key assumption underpinning the successful operation of the Lasso in the high-dimensional setting is that the coefficient vector exhibits sparsity. That is, only a few elements in each $\boldsymbol{\delta}_\ell$ are non-zero.

Condition AS - Approximately Sparse Optimal Instrument: Each optimal instrument function is well-approximated by a function of $s \geq 1$ unknown instruments.

$$\max_{1\le l\le k} ||\boldsymbol{\delta}_l||_0 \le s = o(n) \qquad \max_{1\le l\le k} \left[\frac{1}{n}\sum_i \nu_{i\ell}^2\right]^{1/2} = O_p(\sqrt{s/n}) \tag{47}$$

This condition requires that the true CEF for each instrument exhibits sufficient smoothness to be well-approximated by at most s terms. Define $T_l = \mathit{support}(\boldsymbol{\delta}_l)$. Since s and p are both allowed to grow with n, we also require the following growth condition.

$$\frac{s^2 \log^2(\max\{p, n\})}{n} \to 0 \tag{48}$$

These conditions are analogous to those required for ℓ_2-consistency in section 2.3. An analogue to the restricted eigenvalue condition is also required. First define the restricted set

$$\Delta_{C,T} = \{v \in \mathbb{R}^p : ||v_{T^c}||_1 \le C||v_T||_1, v \neq 0\} \tag{49}$$

for some constant C. The restricted eigenvalue of the Gram matrix, $M = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}$, is

$$\kappa_C^2(M) = {}_{v\in\Delta_{C,T},|T|\le s}\, s\frac{v'Mv}{||v_T||_1^2} \tag{50}$$

which gives the following restricted eigenvalue condition.

Condition RE: For any $C > 0$, there exists a finite constant $\kappa > 0$, which does not depend on n but may depend on C, such that the restricted eigenvalue obeys $\kappa_C(M) \geq \kappa$ with probability approaching 1 as $n \to \infty$.

The first stage regression uses the following version of the Lasso.

$$\hat{\boldsymbol{\delta}}_\ell = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (x_{i\ell} - \boldsymbol{z}_i' \boldsymbol{\delta})^2 + \frac{\lambda}{n} || \hat{\boldsymbol{\Upsilon}}_\ell \boldsymbol{\delta} ||_1 \tag{51}$$

where $\hat{\boldsymbol{\Upsilon}}_\ell = \operatorname{diag}(\hat{\gamma}_{\ell 1}, \ldots, \hat{\gamma}_{\ell p})$ is a matrix of penalty loadings. These loadings allow the Lasso to achieve consistency despite disturbances being non-Gaussian and heteroscedastic using the moderate deviation theory for self-normalizing sums due to Jing et al. (2003). An iterative algorithm for generating these loadings is given in appendix A of Belloni et al. (2012) based on residuals estimated by running a pilot Lasso.

Let $G_{i\ell} = \boldsymbol{z}_i' \boldsymbol{\delta}_\ell$ be the true conditional expectation function and $\hat{G}_{i\ell} = \boldsymbol{z}_i' \hat{\boldsymbol{\delta}}_\ell$ be the Lasso fit. Then we have the following theorem.

Theorem 9 (Belloni et al., 2012)**.** *Suppose Condition AS holds,* $\lambda = 2.2\sqrt{n}\Phi^{-1}\left(1 - \alpha/(2kp)\right)$ *with* $\alpha \to 0$ *and* $\log(1/\alpha) = O(\log(\max\{p, n\}))$*, and* $\hat{\boldsymbol{\Upsilon}}_\ell$ *is generated as above. Then the Lasso fit satisfies*

$$\underset{1 \leq \ell \leq k}{max} \frac{1}{n} ||\hat{G}_{i\ell} - G_{i\ell}||_2 \leq A \frac{1}{\kappa_{\bar{C}}} \sqrt{\frac{s \log(kp/\alpha)}{n}}$$

where $\bar{C}$ is a function of the limiting values of the penalty loadings and κ_C is the corresponding restricted eigenvalue. We omit these values for brevity. Also omitted is an additional condition, called Condition RF, which requires that certain moments of $\boldsymbol{X}$ and $\boldsymbol{Z}$ be bounded.

The upshot of theorem 9 is that we obtain predictive consistency of the Lasso in the first stage with a pivotal choice of λ, despite non-Gaussian heteroscedastic errors. That is, we can consistently estimate the value of the unknown conditional expectation function that yields the optimal instrument.

4.2. Second Stage

With an estimate of the optimal instrument, we may now construct the second stage IV estimator, equation (45), per equation (39). For consistent estimation of $\boldsymbol{\beta}$ we will need the following condition.

Condition SM:

(i) The eigenvalues of $\mathbb{E}[G(\boldsymbol{Z})G(\boldsymbol{Z})']$ are bounded above and away from zero uniformly in n.

(ii) For some $q > 2$ and $q_\varepsilon > 2$, uniformly in n,

$$\max_{1\le j\le p} \mathbb{E}[|z_{ij}\varepsilon_j|^3]+\mathbb{E}[||G_i||_2^q|\varepsilon_i|^{2q}]+\mathbb{E}[||G_i||_2^q]+\mathbb{E}[|\varepsilon_i|^{q_\varepsilon}]+\mathbb{E}[||\boldsymbol{x}_i||_2^q] = O(1)$$

(iii) The following growth conditions hold.

(a) $\log^3 p = o(n)$

(b) $\frac{s\log(\max\{p,n\})}{n} n^{2/q_\varepsilon} \to 0$

(c) $\frac{s^2\log^2(\max\{p,n\})}{n} \to 0$

(d) $\max_{1\le j\le p} \frac{1}{n}\sum z_{ij}^2\varepsilon_i^2 = O_p(1)$

The first condition is a strong identification assumption requiring that we do not have weak instruments, the second is a mild moment assumption, and the third requires that the dimension and number of relevant regressors grow much more slowly that n; more slowly than the usual rates required for the Lasso.

Now we may state the key theorem.

Theorem 10 (Belloni et al., 2012)**.** *Suppose the data obey the linear model above. Suppose conditions AS, RF, RE, and RM hold with $\lambda = 2.2\sqrt{n}\Phi^{-1}\left(1-\alpha/(2kp)\right)$. Then the IV estimator, $\hat{\boldsymbol{\beta}}_{IV}$ based on Lasso estimates of the optimal instrument is root-n consistent and asymptotically normal.*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{IV}-\boldsymbol{\beta}\right) \xrightarrow{d} N(0, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}'^{-1})$$

where $\boldsymbol{\Omega} = \mathbb{E}[\varepsilon_i^2 G(\boldsymbol{Z})G(\boldsymbol{Z})']$ and $\boldsymbol{Q} = \mathbb{E}[G(\boldsymbol{Z})G(\boldsymbol{Z})']$.

Moreover, the asymptotic variance may be replaced with a consistent estimator and in the event that disturbances are conditionally heteroscedastic, $\mathbb{E}[\varepsilon_i^2\,|\boldsymbol{Z}] = \sigma^2$, the estimator attains the semiparametric efficiency bound as in equation (41).

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}\right) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{Q}^{-1})$$

Thus inference may proceed as usual when strong identification obtains. Belloni et al. (2012) further develop a sup-score test, which remains valid in the face of weak instruments which may be inverted to produce an identification-robust confidence set, and a Hausmann-type specification test for the validity of a subset of instruments.

5. Discussion

The Lasso is a powerful object in an applied researcher's toolbox. It performs the simultaneous feats of model selection and parameter estimation. While it has been a staple in the statistics and machine learning communities for at least the previous 15 years, it has yet to attract popularity in applied economic research. This could potentially be due to the narrow focus of econometrics on identification and causal inference. Statistical and machine learning problems often tend to be purely predictive, such as deciding if an email is spam based on the words it contains or classifying a handwritten number based on a photograph. For these types of problems the Lasso has a very clear advantage over ordinary least squares, which is that the Lasso produces the best sparse linear predictor (Greenshtein and Ritov, 2004).

For problems of causal inference OLS has a clear advantage, its asymptotic distribution is easy to characterize and it is unbiased, meaning that confidence intervals are easy to construct. The Lasso on the other hand is a biased estimator for which the asymptotic distribution has no closed form. In low dimensions the Lasso may be bootstrapped, though with considerably more difficulty than the OLS estimator, and in high dimensions we may resort to sample splitting, which some researchers may find unpalatable. However, there are good arguments for choosing the Lasso even if the objective is causal inference. The validity of inferences based on OLS come with two important caveats.

First, the true regression function must be linear. Since it is widely understood that this assumption is rather unlikely to hold, the linear model is usually interpreted as a local linear approximation to the true CEF. The Lasso can easily be used to model non-linearities in the CEF by including transformations and interactions of the regressors.

Second, the p-values and confidence intervals only have valid size and coverage if the model reported was the only model estimated by the researcher. If a researcher makes decisions about which variables enter a model based on running regressions on various combinations of variables, this constitutes a form of model selection and inferences based on the resulting model will fail to obtain their advertized frequency guarantees. That is, 95% confidence intervals will have less than 95% coverage. The Lasso automates model selection and internalizes the selection process when conducting inference.

A particularly auspicious application of the Lasso in applied economics is to instrumental variables regression. A prominent issue faced by IV regression is the many instruments problem. Adding many instruments to a model, which intuitively should improve the performance of the estimator, can actually make the variance of the IV estimator unboundedly large. The problem is to find the right combination of a few instruments which predict the endogenous variable well. The Lasso has the very valuable property that it is a *sparse* estimator. It can find the best linear combination of instruments that predict the endogenous variable in the first stage while controlling the variance of the resulting IV estimator in the second stage. It does this by setting the weights on some of the instruments *exactly* to zero, effectively removing them from the model. Were the Lasso estimator to enter mainstream econometric practice, this setting would likely be its beachhead.

Amemiya, T. (1966). On the use of principal components of independent variables in two-stage least-squares estimation. *International Economic Review*, 7(3):283–303.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6):1341–1383.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of Economic Perspectives*, 24(2):3–30.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79.

Bai, J. and Ng, S. (2009). Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1):1941–1928.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2009). Least squares after model selection in high-dimensional sparse models.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791.

Bhlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6):1196–1217.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*.

Caner, M. (2009). Lasso-type gmm estimator. *Econometric Theory*, 25(1):270–290.

Chamberlain, G. and Imbens, G. (2004). Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306.

Chang, A. C. and Li, P. (2015). Is economics research replicable? sixty published papers from thirteen journals say'usually not'. *http://dx.doi.org/10.17016/FEDS.2015.083.*

Chatterjee, A. and Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

Dezeure, R., Bhlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and r-software hdi. *Statist. Sci.*, 30(4):533–558.

Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69(5):1161–1191.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.*, 9(6):1218–1228.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3):1–15.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8).

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

Jing, B.-Y., Shao, Q.-M., and Wang, Q. (2003). Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215.

Kloek, T. and Mennes, L. B. M. (1960). Simultaneous equations estimation based on principal components of predetermined variables. *Econometrica*, 28(1):45–61.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2):413–468.

Meinshausen, N. (2015). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):923–945.

Meinshausen, N., Meier, L., and Bhlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 165(1):70 – 86. Moment Restriction-Based Econometric Methods.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, 37(5A):2178–2201.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.