

Applications of Machine Learning on Econometrics for Two-stage Regression, Bias-adjusted Inference with Unobserved Confounding, and Test for High Dimensionality

by

Wenzuo Xu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Economics

Waterloo, Ontario, Canada, 2024

© Wenzuo Xu 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor: Tao Chen
Associate Professor
Department of Economics
University of Waterloo

Internal Members: Pierre Chaussé
Associate Professor
Department of Economics
University of Waterloo

Thomas Parker
Associate Professor
Department of Economics
University of Waterloo

Internal-External Member: Pengfei Li
Professor
Department of Statistics and Actuarial Science
University of Waterloo

External Member: Xin Jin
Associate Professor
Department of Economics
Shanghai University of Finance and Economics

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

While Chapter 1, 2 and 3 are not sole-authored (Chapter 1 is co-authored with my supervisor Professor Tao Chen and Assistant Professor Renfang Tian from the school of management at the King's University College; Chapter 2 is co-authored with my supervisor Professor Tao Chen and Assistant Professor Renfang Tian from the school of management at the King's University College; Chapter 3 is co-authored with my supervisor Professor Tao Chen), I have made the major contribution to the work involved in proving the results.

For Chapter 1, I have made the major contribution to most of the proofs and the simulation study. For Chapter 2, the topic is extended by my term project in ECON 723 and I have made the contribution to all the proofs and the simulation study. For Chapter 3, I have made the contribution to all the proofs and the simulation study.

Abstract

Nonparametric approaches have been extensively studied and applied when no assumption is made regarding the model specification. Estimation of classical nonparametric regression models may involve methods such as spline regression, local polynomial regression, and kernel regression, etc. More generally, a sieve can be constructed as a collection of subsets of finite-dimensional approximating parameter spaces, over which the target function is estimated by an optimization of fitting without demanding a parametric specification (e.g., Grenander, 1981; Chen, 2007). Becoming sufficiently rich and dense in the whole space with an enlarging sample, the sieves induce well-posed optimization problems and allow consistent estimation of both parametric and nonparametric components with optimal convergence rates (e.g., Grenander, 1981; Chen, 2007; Chen and Liao, 2015).

Although the concept of sieves is devised in such a general way, classic sieve estimation in literature has been mostly focusing on “single-layer” approximations. When the target functions are of intricate patterns, however, these single-layer estimators show limited capability despite allowance for data-generated sieve bases (e.g., free-knot splines or trained neural networks), whereas characterizing different attributes of the target functions progressively through multiple layers is often more sensible (e.g., Fabozzi et al., 2019; Mathew et al., 2021).

Deep neural networks (DNNs) offer a “multi-layer” extension of the traditional sieves by modelling the connections among variables through data transformations from one layer to another (e.g., Fabozzi et al., 2019; Shen, Xiaoxi and Jiang, Chang and Sakhanenko, Lyudmila and Lu, Qing, 2019; Horel and Giesecke, 2020; Farrell et al., 2021). A major concern about DNNs has been raised that they are mostly applied and treated as black boxes with a lack of interpretability, in the sense that the approximating networks do not provide any intuition on the structure of the functions being approximated. Therefore, we introduce the polynomial DNN structure for the two stage estimation in Chapter 1. While preserving the basic multi-layer network structure and the capability of DNNs, the polynomial activation functions can help crack open the black boxes. Due to the nice property that the composition of polynomials is still a polynomial, using polynomial activation functions in a DNN essentially generates a polynomial approximation, which allows for interpretation of the estimated functions as polynomials. The estimated parameters of

the network will be involved in the polynomial coefficients that determine the explanatory power of the monomial terms, and the approximating network can also illustrate how each input relates to the outputs. By improving the interpretability of the network, we can gain a better understanding of the approximated structure and the role plays in the first stage.

DNNs have a larger freedom than the single-layer ones in increasing the sieve complexity to ensure consistent estimation while maintaining a relatively simple structure in each layer for feasible estimation. Although Chen et al. (2023) demonstrate via simulation and empirical studies that, in certain situations (e.g., estimating the density-weighted average derivative in a nonparametric/semiparametric model), multi-layer networks have no clear advantage over some single-layer sieves in terms of efficiency, Mathew et al. (2021), on the other hand, state that the performance of deep learning improves on a larger scale with a growing sample size compared to traditional single-layer learning methods. Taking advantage of such properties, we propose DNN-based estimators in Chapter 2 to estimate the underlying relationship for the treatment in a semi-parametric scenario with unobserved confounding.

DNNs have gradually evolved into state-of-the-art technology with a wide range of tasks in practical such as image classification and restoration (e.g., Rawat and Wang, 2017; Dong et al., 2018), speech recognition (Fohr et al., 2017), game intelligence (Perolat et al., 2022), and natural language processing (Conneau et al., 2016). The main scenarios to be investigated steadily evolve into the “large dimension and large sample size”. Although many methods have been proposed for high-dimensional scenario; however, a more basic question about whether random vectors in a finite data sample is in high dimensionality or not, has not been considered in the literature. To deal with such problems, we proposes a general testing method to distinguish high dimensionality of random vectors from non-high in Chapter 3.

Acknowledgments

I wish to thank Professors Tao Chen, Pierre Chaussé, Thomas Parker and Pengfei Li for their sound advice and kind support. Without them, this thesis would not have been possible. Special thanks to my supervisor, Tao Chen. I have learned from him that every milestone is a new start, and life is a process of endless exploration and discovery. I would also like to express my gratitude to all the members of faculty, staff and my colleagues, who have helped me and offered deep insight into my PhD. Finally, I thank my parents and my friends for their constant support and encouragement.

Dedication

To my parents I dedicate this thesis.

Table of Contents

| | |
|---|----------|
| List of Figures | xix |
| List of Tables | xxi |
| Introduction | 1 |
| 1 Deep Neural Network Empowered Two-stage Regression with a Non-linear First Stage | 3 |
| 1.1 Introduction | 3 |
| 1.2 DNN-based Estimator and its Large Sample Properties | 6 |
| 1.2.1 DNN-based estimators | 7 |
| 1.2.2 Large sample properties | 9 |
| 1.2.3 Generalized parametric structure model | 13 |
| 1.3 Tests for Weak Identification | 15 |
| 1.4 Simulation Analysis | 18 |
| 1.4.1 Global Settings | 19 |
| 1.4.2 Simulation Designs | 19 |
| 1.4.3 Estimation | 21 |
| 1.4.4 Results and discussion | 22 |
| 1.5 Conclusion | 31 |

| | | |
|----------|--|-----------|
| 2 | Bias-adjusted Inference with Unobserved Confounding | 33 |
| 2.1 | Introduction | 33 |
| 2.2 | Methodology and Asymptotics | 36 |
| 2.2.1 | General Setup and Definitions | 36 |
| 2.2.2 | Scenario 1: Fully represented unobserved part | 37 |
| 2.2.3 | Scenario 2: Partially represented unobserved part | 42 |
| 2.3 | Simulation Analysis | 45 |
| 2.3.1 | Simulation Study for Scenario 1 | 46 |
| 2.3.2 | Simulation Study for Scenario 2 | 47 |
| 2.4 | Conclusion | 51 |
| 3 | Test for High Dimensionality of Random Vectors | 53 |
| 3.1 | Introduction | 53 |
| 3.2 | Tests of Random Vectors | 56 |
| 3.2.1 | Multivariate Normal Vectors | 56 |
| 3.2.2 | General Random Vectors | 57 |
| 3.3 | Implementation | 60 |
| 3.3.1 | Multivariate normal random vectors with identity covariance matrix | 61 |
| 3.3.2 | Multivariate normal random vectors with non-identity covariance matrix | 61 |
| 3.3.3 | General random vectors with known expected mean | 62 |
| 3.3.4 | General random vectors with unknown expected mean | 63 |
| 3.4 | Monte Carlo Study | 63 |
| 3.4.1 | Simulation Designs | 64 |
| 3.4.2 | Results and discussion | 66 |
| 3.5 | Conclusion | 75 |

| | |
|--|-----------|
| References | 77 |
| APPENDICES | 87 |
| A Appendices of Chapter 1 | 87 |
| A.1 Proof of Theorems | 88 |
| A.2 Proof of Lemmas | 95 |
| B Appendices of Chapter 2 | 98 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Deep Neural Network | 8 |
| 1.2 | Distributions of the OLS, 2SLS and DNN in Design (i) | 23 |
| 1.3 | Distributions of the OLS, LIML, Sieve and DNN in Design (i) | 25 |
| 1.4 | Distributions of the OLS, 2SLS and DNN Estimators in Design (ii) | 27 |
| 1.5 | Distributions of the DNN Estimators in Design (iii) | 29 |
| 1.6 | Distributions of the OLS, 2SLS and DNN in Design (iv) | 30 |
| 2.1 | Deep Neural Network architecture with ReLU activation | 38 |
| 2.2 | Distributions of the OLS and DNN Estimator in Scenario 1, $n = 1000$ | 48 |
| 2.3 | Distributions of the OLS and DNN Estimator in Scenario 1, $n = 2000$ | 48 |
| 2.4 | Distributions of the OLS and DNN Estimator in Scenario 2, $n = 1000$ | 50 |
| 2.5 | Distributions of the OLS and DNN Estimator in Scenario 2, $n = 2000$ | 51 |

List of Tables

| | | |
|------|--|----|
| 1.1 | Estimation Statistics for OLS, 2SLS and DNN in Design (i) | 23 |
| 1.2 | Estimation Statistics for OLS, LIML, Sieve and DNN in Design (i) | 24 |
| 1.3 | Estimation Statistics for OLS, 2SLS and DNN in Design (ii) | 26 |
| 1.4 | Estimation Statistics for OLS and DNN estimators in Design (iii) | 27 |
| 1.5 | Estimation Statistics for OLS, 2SLS, and DNN in Design (iv) | 28 |
| 2.1 | Scenario 1: Estimation Statistics | 47 |
| 2.2 | Scenario 2: Estimation Statistics | 49 |
| 3.1 | Empirical rejection rates in Designs (i) | 69 |
| 3.2 | Empirical rejection rates in Designs (ii) | 70 |
| 3.3 | Empirical rejection rates in Designs (iii) | 71 |
| 3.4 | Empirical rejection rates in Designs (iv) | 72 |
| 3.5 | Empirical rejection rates in Designs (v) | 73 |
| 3.6 | Dividing Line D^* and D^{**} with N in Design (i) | 74 |
| 3.7 | Dividing Line D^* and D^{**} with N in Design (ii) | 74 |
| 3.8 | Dividing Line D^* and D^{**} with N in Design (iii) | 74 |
| 3.9 | Dividing Line D^* and D^{**} with N in Design (iv) | 74 |
| 3.10 | Dividing Line D^* and D^{**} with N in Design (v) | 74 |

Introduction

This thesis contains three chapters developing methodologies and motivating applications of machine learning on Econometrics for two-stage regression, bias-adjusted inference with unobserved confounding, and test for high dimension.

In Chapter 1, co-authored with Tao Chen and Renfang Tian, we incorporate DNN in a two-stage regression model to uncover the relationship in the first stage. The idea is motivated by the fact that a linear first stage is not guaranteed to correctly specify the relationship between the instrumental variables and the endogenous explanatory variables in causal inference with traditional two-stage least squares regression. Such misspecification may weaken the model identification and lead to unreliable inference subsequently. We incorporate DNNs in a two-stage regression, where the true first-stage relationship is unknown and may be non-linear. This study contributes to literature in two main aspects. First, we develop a DNN-based estimator formulated on the polynomial activation functions and derive its large sample properties. By improving the interpretability of the DNN network structure, we can gain a better understanding of the approximated structure and the role each instrument variable plays. Second, we construct a generalized first-stage F-test for weak IVs and verify the test validity. Since weak identification leads to unreliable causal inference, it is essential to test the strength of the IVs before drawing any conclusions. In practice, if a traditional F-test shows that the IVs are weak under predetermined parametric specifications, one can further test for weak identification under the DNN fitting by using this generalized first-stage F-test — if the IVs turn out to be strong enough, our DNN-based estimator can be applied, while if the IVs are still shown to be weak, we need to switch to better IVs.

In Chapter 2, co-authored with Tao Chen and Renfang Tian, we develop a method for reliable bias-adjusted inferences in two scenarios – when the unobserved confounder can be fully explained by the observed controls and when it cannot be fully explained. The motivation starts from the fact that the omitted variable bias constitutes a major concern in causal inference, whence control variables are included to capture the omitted factors. Yet, available controls often fail to provide complete proxies for the unobserved confounder. Previous studies have discussed issues such as reducing the estimation bias, estimating the bounds of the bias, and deriving the linkage between the bias and the

movements of the statistics, etc. However, bias-adjusted inferences were often conducted under restrictive conditions that barely approaches the practice. We propose DNN-based estimators in both two scenarios. We prove the consistency of the DNN-based estimators in both scenarios and derive asymptotic normality for the estimator from scenario one, while the inference for scenario two is performed by bootstrap. Essentially, the proposed method addresses the issue of omitted variable bias in observational studies by isolating the effects from unobserved confounding using DNN. To our best knowledge, such a method has not been developed in previous studies.

In Chapter 3, co-authored with Tao Chen, we proposes a general testing method to distinguish high dimensionality of random vectors from non-high . Prior to 1950, most of practical problems consisted of a relatively large number of experimental units with a relatively small number of features (Rowell and Walters, 1976). Therefore, traditional theories and practice were limited to the “small dimension of variables and large sample size” scenario. Over the last 25 years, however, Lindsay et al. (2004) pointed out that the main scenarios to be investigated steadily evolve into the “large dimension and large sample size”. Although many methods have been proposed for high-dimensional scenario; however, a more basic question about whether random vectors in a finite data sample is in high dimensionality or not, has not been considered in the literature. The idea behind our proposed test for high dimension is comparable to that behind the central limit theorem (CLT) in which the normal distribution appears in the case of a sufficiently large sample size. Similar to determining how large the sample size N is adequate to hold the CLT in a finite sample, one might ask how large the dimension D of random vectors need to be for these high-dimensional theories to hold. We propose our test for high dimensionality of random vectors and provide guidance to determine the thresholds in the test. The simulation study shows the size and power of our proposed test and illustrates that if we observed enough number of random vectors, only the underlying distribution of the random vectors in the finite sample alters the dimension that is needed for high dimensionality. If the underlying distribution is more complex, a larger dimension of random vectors is required for high dimensionality as the cost to achieve the high dimension properties.

Chapter 1

Deep Neural Network Empowered Two-stage Regression with a Non-linear First Stage¹

1.1 Introduction

Two-Stage least squares (2SLS) regression analysis with instrumental variables (IVs) provides a classic solution when endogeneity in the explanatory variables is suspected (e.g., Angrist et al., 1996; Heckman, 1997; Heckman and Vytlacil, 2001; Lee, 2003; Heckman and Navarro-Lozano, 2004). In traditional 2SLS regression with a linearly-specified first stage, not only need the IVs be exogenous, they also need to have a sufficiently strong linear relationship with the endogenous explanatory variables to ensure the validity of the results. If the linear first stage can only capture a weak relationship, the model will encounter the problem of weak identification (e.g., Staiger and Stock, 1997; Stock et al., 2002; Hansen et al., 2008; Andrews et al., 2019). Such a problem is fixable if the identification can be enhanced by proper first-stage modelling. A worse case scenario is that the IVs are not relevant enough in any format, and this is when better IVs are needed.

The current chapter tackles the cases where a linear first stage only induces weak

¹This chapter is co-authored with Tao Chen and Renfang Tian.

identification, yet the true relationship is strong enough but in an unknown non-linear format. In such circumstances, a more general parametric first stage is also sub-optimal since it is not feasible to exhaust all possible parametric specifications until a relationship found is strong enough to provide valid causal inference. Therefore, methods that offer reliable fitting without demanding the functional form naturally become more appropriate.

Nonparametric approaches have been extensively studied and applied when no assumption is made regarding the model specification. Estimation of classical nonparametric regression models may involve methods such as spline regression, local polynomial regression, and kernel regression, etc. More generally, a sieve can be constructed as a collection of subsets of finite-dimensional approximating parameter spaces, over which the target function is estimated by an optimization of fitting without demanding a parametric specification (e.g., Grenander, 1981; Chen, 2007). Becoming sufficiently rich and dense in the whole space with an enlarging sample, the sieves induce well-posed optimization problems and allow consistent estimation of both parametric and nonparametric components with optimal convergence rates (e.g., Grenander, 1981; Chen, 2007; Chen and Liao, 2015).

Although the concept of sieves is devised in such a general way, classic sieve estimation in literature has been mostly focusing on “single-layer” approximations. Typical examples include splines, wavelets, power series, Fourier series, kernel, and single-layer neural networks (NNs) (e.g., De Boor and De Boor, 1978; Gallant, 1988; Chen and Shen, 1998; Chen, 2007; Schumaker, 2007). When the target functions are of intricate patterns, however, these single-layer estimators show limited capability despite allowance for (partially) data-generated sieve bases (e.g., free-knot splines or trained neural networks), whereas characterizing different attributes of the target functions progressively through multiple layers is often more sensible (e.g., Fabozzi et al., 2019; Mathew et al., 2021).

Deep neural networks (DNNs) offer a “multi-layer” extension of the traditional sieves by modelling the connections among variables through data transformations from one layer to another (e.g., Fabozzi et al., 2019; Shen, Xiaoxi and Jiang, Chang and Sakhanenko, Lyudmila and Lu, Qing, 2019; Horel and Giesecke, 2020; Farrell et al., 2021). Such networks have a larger freedom than the single-layer ones in increasing the sieve complexity to ensure consistent estimation while maintaining a relatively simple structure in each layer for feasible estimation. Although Chen et al. (2023) demonstrate via simulation and empirical studies that, in certain situations (e.g., estimating the density-weighted average derivative

in a nonparametric/semiparametric model), multi-layer networks have no clear advantage over some single-layer sieves in terms of efficiency, Mathew et al. (2021), on the other hand, state that the performance of deep learning improves on a larger scale with a growing sample size compared to traditional single-layer learning methods. Recent studies have suggested to use NNs to extract the relationship between the endogenous explanatory variables and the IVs (e.g., Hartford et al., 2017; Chen et al., 2020; Liu et al., 2020; Sheu, 2020; Farrell et al., 2021). Chen et al. (2020) have also specifically pointed out that incorporating a DNN can dramatically improve estimation precision and robustness by boosting IVs' strength.

In the current chapter, we incorporate DNNs in a two-stage regression, where the true first-stage relationship is unknown and may be non-linear. This study contributes to literature in two main aspects. First, we develop a DNN-based estimator formulated on the polynomial activation functions and derive its large sample properties. There has been literature deriving DNN-based estimators (e.g., Chen et al., 2020; Liu et al., 2020) under different activation functions (such as ReLU). However, a major concern about DNNs has been raised that they are mostly applied and treated as black boxes with a lack of interpretability, in the sense that the approximating networks do not provide any intuition on the structure of the functions being approximated (e.g., Emmons et al., 2019; Sheu, 2020), and this is due to the multi-layer non-linear structures under various activation functions (Buhrmester et al., 2021). Such sophistication offers capability and flexibility in fitting complex data structures yet also makes the approximating network itself less intuitive. While preserving the basic multi-layer network structure and the capability of DNNs, the polynomial activation functions can help crack open the black boxes. Due to the nice property that the composition of polynomials is still a polynomial, using polynomial activation functions in a DNN essentially generates a polynomial approximation, which allows for interpretation of the estimated functions as polynomials. The estimated parameters of the network will be involved in the polynomial coefficients that determine the explanatory power of the monomial terms, and the approximating network can also illustrate how each input relates to the outputs. By improving the interpretability of the network, we can gain a better understanding of the approximated structure and the role each IV plays.

Second, we construct a generalized first-stage F-test for weak IVs and verify the test validity. Since weak identification leads to unreliable causal inference, it is essential to test the strength of the IVs before drawing any conclusions. The first-stage F-statistic has

been commonly used in traditional 2SLS regression to test for weak IVs with the rule-of-thumb criterion $F < 10$ (from Stock and Yogo (2002), to which if referred correctly, should have been $F < 104.7$ as pointed out by Lee et al. (2021)), where the F-statistic measures the relative size of the explained versus the unexplained part of the first stage. Our estimator employs DNNs for the first-stage fitting, whence the traditional F-statistic is no longer adequate, but a similar concept can be adopted. In this chapter, we construct a test statistic based on an established measurement for the instruments' overall explanatory power — the concentration parameter — and obtain the relative size of the signal versus the noise under the DNN fitting, where the polynomial network structure also helps simplify the verification of the test validity. In practice, if a traditional F-test shows that the IVs are weak under predetermined parametric specifications, one can further test for weak identification under the DNN fitting by using this generalized first-stage F-test — if the IVs turn out to be strong enough, our DNN-based estimator can be applied, while if the IVs are still shown to be weak, we need to switch to better IVs.

The rest of this chapter is structured as follows. In Section 1.2, we introduce our DNN-based estimator and establish its large sample properties. In Section 1.3, we construct our test statistic and interpret its validity. In section 1.4, we illustrate the proposed method in a simulation analysis. Section 1.5 concludes.

1.2 DNN-based Estimator and its Large Sample Properties

Consider the following structural model

$$Y = \beta_0^\top X + U, \tag{1.2.1}$$

with a first stage

$$X = g(Z) + V. \tag{1.2.2}$$

Here, Y denotes the scalar response, X a d_X -vector of non-degenerate explanatory variables, β_0 a d_X -vector of coefficients, U and V the random errors, Z a d_Z -vector ($d_Z \geq d_X$) of non-degenerated IVs that are continuously distributed on $[0, 1]^{d_Z}$ and exogenous in that

$\mathbb{E}[U|Z] = \mathbb{E}[V|Z] = 0$, and $g : [0, 1]^{d_Z} \rightarrow \mathbb{R}^{d_X}$ a well-defined non-stochastic continuous matrix function of an unknown form. Note that the proposed estimator is motivated under the linear structural model (1.2.1). The interpretation of this estimator under a generalized parametric structural model is discussed below in Section 1.2.3.

Due to its continuity, each entry of g , denoted by $g_r : [0, 1]^{d_Z} \rightarrow \mathbb{R}$ for $r = 1, \dots, d_X$, has an arbitrarily close polynomial approximation according to the Weierstrass approximation theorem, and the multivariate Bernstein polynomial $B_K(z_1, \dots, z_{d_Z}, g_r)$ provides a specific format of polynomial approximation, as such for each $K := [K_1, \dots, K_{d_Z}] \in \mathbb{N}^{d_Z}$,

$$B_K(z_1, \dots, z_{d_Z}, g_r) := \sum_{\substack{0 \leq k_m \leq K_m \\ m \in \{1, \dots, d_Z\}}} g_r \left(\frac{k_1}{K_1}, \dots, \frac{k_{d_Z}}{K_{d_Z}} \right) \prod_{m=1}^{d_Z} \left\{ \binom{K_m}{k_m} z_m^{k_m} (1 - z_m)^{K_m - k_m} \right\}. \quad (1.2.3)$$

Lemma 1.2.1. *Consider $B_K(z_1, \dots, z_{d_Z}, g_r)$ as in (1.2.3) for the continuous underlying function g_r with finite d_Z . For any $\varepsilon > 0$, there exists some $\delta > 0$ such that $\|z^* - z^0\| \leq \delta \implies |g_r(z^*) - g_r(z^0)| \leq \varepsilon/2$ for all $z^*, z^0 \in \mathbb{R}^{d_Z}$. Then with equal order \bar{K} for all z_1 to z_{d_Z} , say $\bar{K} := K_1 = \dots = K_{d_Z}$, $\bar{K} \geq (d_Z c)/(\delta^2 \varepsilon)$ for $c := \sup_{z \in [0, 1]^{d_Z}} |g_r(z)| < \infty$ implies that*

$$\sup_{z \in \mathbb{R}^{d_Z}} |B_K(z, g_r) - g_r(z)| \leq \varepsilon.$$

Lemma 1.2.1 suggests that, fixing everything else, for any given ε , one needs a high-degree polynomial approximation if the continuity of g_r requires a small δ , while low-degree polynomials can be used to achieve a desired approximation if a large δ suffices the given ε . Under such multivariate Bernstein polynomial approximations $B_K(z_1, \dots, z_{d_Z}, g_r)$ for all $r = 1, \dots, d_X$, the first-stage functional form g can be arbitrarily closely represented by a summation as such $g(Z) \approx X_l(Z) + X_{nl}(Z)$, where $X_l(Z) := \pi_0^\top Z$ is a linear component with a d_Z -by- d_X real-valued coefficients matrix π_0 and $X_{nl}(Z) : \mathbb{R}^{d_Z} \rightarrow \mathbb{R}^{d_X}$ an unknown non-linear component.

1.2.1 DNN-based estimators

To estimate the function g , we represent its Bernstein polynomial approximation with a DNN as shown in Figure 1.1. The network contains L hidden layers and a set of corre-

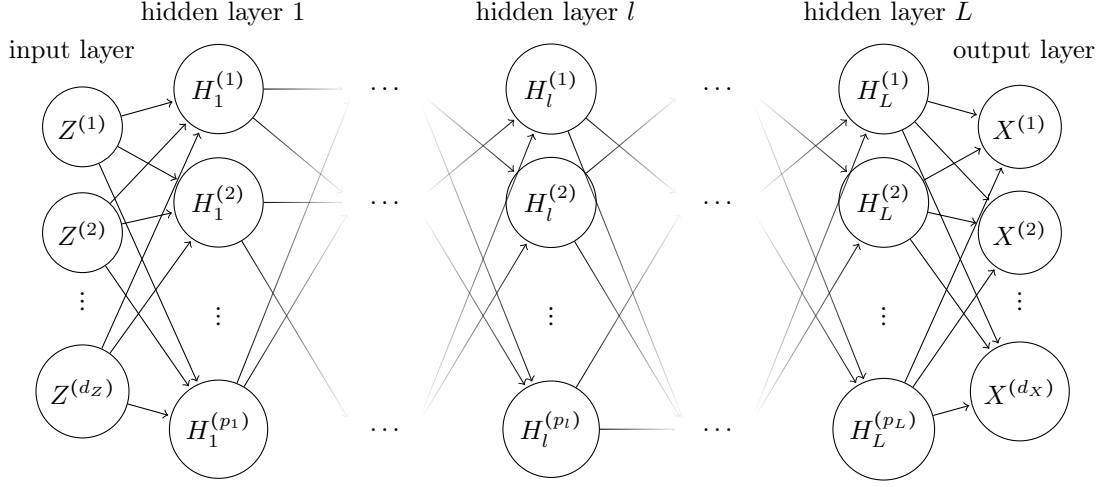


Figure 1.1: Deep Neural Network

spending width in each hidden layer $\mathbf{p} := (p_0, p_1, \dots, p_L, p_{L+1})$, which is a function of the form

$$s(Z_i) = W_L \circ \text{poly} \circ \dots \circ W_2 \circ \text{poly} \circ W_1 \circ \text{poly} \circ W_0 Z_i, \quad (1.2.4)$$

where p_l denotes the number of units in the l -th hidden layer for $l = 1, \dots, L$. p_0 and p_{L+1} denotes the number of inputs and outcomes, where $p_0 = d_Z$ and $p_{L+1} = d_X$. We define the set of weight matrices $W = \{W_0, \dots, W_L\}$, where W_l projects hidden layer l to $l + 1$ for $l = 1, 2, \dots, L - 1$, W_0 projects the input layer to the first hidden layer, and W_L projects the last hidden layer to the outcome layer. Each W_l is a matrix of $p_{l+1} \times p_l$ weights, for $l = 0, 1, \dots, L$. The space \mathcal{S} of all such networks is defined as

$$\mathcal{S}(L, \mathbf{p}) := \left\{ s \text{ takes the form of (1.2.4)} : \max_{l=0, \dots, L+1} \|W_l\|_\infty \leq \infty \right\},$$

where $\|W_l\|_\infty := \max_{1 \leq i \leq d_{W_l, l+1}} \sum_{j=1}^{d_{W_l, l}} |W_{l, [i, j]}|$.

Assumption 1.2.1. We have a size- n random sample $(Y_i, X_i, Z_i)_{i=1}^n$ of the variables (Y, X, Z) .

As indicated by Assumption 1.2.1, we will work with i.i.d. (independent and identically distributed) samples in this chapter, which is a trivial but important special case of those

with independent but not necessarily identically distributed samples or those with stationary and ergodic samples. With proper conditions on some higher moments of the joint distribution of (X_i, Z_i, U_i, V_i) , heteroskedasticity or weak dependence among individuals can also be accommodated.

We denote a DNN representation for g by $g_n^s(Z_i; W)$, with the sample size n , the network structure $s \in \mathcal{S}$, and weights W . Then the estimators of the weights, denoted by $\hat{W} := \{\hat{W}_0, \hat{W}_1, \dots, \hat{W}_L\}$, and thus, the estimated function $g_n^s(Z_i; \hat{W})$ can be obtained by minimizing the loss function

$$\text{Loss}_n^s(\tilde{W}) := \frac{1}{n} \sum_{i=1}^n \|X_i - g_n^s(Z_i; \tilde{W})\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a rectangular matrix, and \tilde{W} represents any estimator of the weights. With the estimated function $g_n^s(Z_i; \hat{W})$, we can define a DNN-based estimator $\hat{\beta}_{\text{DNN}}$ for β_0 , such that

$$\hat{\beta}_{\text{DNN}} := \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) Y_i \right\}. \quad (1.2.5)$$

1.2.2 Large sample properties

We establish the asymptotics of this DNN-based estimator under the following assumptions.

Assumption 1.2.2. *The IVs are exogenous as such $\mathbb{E}[U_i|Z_i] = \mathbb{E}[V_i|Z_i] = 0$ for $i = 1, \dots, n$.*

Assumption 1.2.2 is sufficient to ensure the exogeneity under any form of the first-stage relationship g , yet it is not a necessary exogeneity condition. As discussed below, if polynomial activation functions are adopted in the DNN, the estimated function $g_n^s(Z_i; \hat{W})$ will also be a polynomial, and thus, is linear in parameters. Then, among other necessary assumptions, a proper high-order cross-moment condition $\mathbb{E}[U_i P(Z_i)] = 0$ for $i = 1, \dots, n$, where $P(Z_i)$ represents a vector of power functions of the IVs (further discussed in the appendix), also implies desired consistency and asymptotic normality of the estimator $\hat{\beta}_{\text{DNN}}$.

Assumption 1.2.3. *For all $i = 1, \dots, n$*

(a). there exists a strictly positive-definite real-valued matrix Σ_{UV} , such that

$$\Sigma_{UV} := \begin{bmatrix} \sigma_{UU} & \sigma_{UV} \\ \sigma_{UV}^\top & \sigma_{VV} \end{bmatrix} := \text{Var} \left(\begin{bmatrix} U_i \\ V_i \end{bmatrix} \middle| Z_i \right).$$

(b). there exist real-valued matrices $\Sigma_{gX} := \mathbb{E}[g(Z_i)X_i^\top]$ and $\Sigma_{BX} := \mathbb{E}[B_K(Z_i, g)X_i^\top]$ for $B_K(Z_i, g) := [B_K(Z_i, g_1) \cdots B_K(Z_i, g_{d_X})]^\top$, such that Σ_{gX} and Σ_{BX} have full rank for all n , for at least all but finitely many \bar{K} .

Assumption 1.2.3(a) imposes regularization on the cross moments among the error terms, and conditional homoskedasticity is assumed as stated previously. Part (b) ensures the invertibility of Σ_{gX} and Σ_{BX} , which indicates that the IVs are strong enough under the true relationship g and its multivariate Bernstein polynomial approximations $B_K(Z_i, g)$.

Consistency

Having that the first-stage underlying function g can be uniformly approximated by multivariate Bernstein polynomials, which in turn can be estimated by the DNN $g_n^s(Z_i; \hat{W})$, the following result holds.

Lemma 1.2.2. *By the first stage (1.2.2), Assumptions 1.2.1 to 1.2.3, and Lemma 1.2.1, there exists some network s such that $\mathbb{E}[\|g_n^s(z; \hat{W}) - g(z)\|_F^2] \rightarrow 0$ for all $z \in [0, 1]^{d_Z}$ as $n, L, \bar{K} \rightarrow \infty$.*

The error from $g(z)$ to $g_n^s(z; \hat{W})$ consists of two parts — the approximation error ε from $g(z)$ to $B_K(z, g)$ and the estimation error from $B_K(z, g)$ to $g_n^s(z; \hat{W})$. The convergence of the approximation error ε can be achieved through $\bar{K} \rightarrow \infty$ as indicated by Lemma 1.2.1, and the convergence of the estimation error $g_n^s(z; \hat{W}) - B_K(z, g)$ holds as $n, L \rightarrow \infty$. With the convergence of the two error components, the result in Lemma 1.2.2 follows, which implies the consistency of $\hat{\beta}_{\text{DNN}}$.

Theorem 1.2.1. *Suppose Assumptions 1.2.1 to 1.2.3 hold. With \bar{K} satisfying Assumption 1.2.3(b), $g_n^s(Z_i; \hat{W})$ satisfying Lemma 1.2.2, and $\hat{\beta}_{\text{DNN}}$ as in (1.2.5), there is $\|\hat{\beta}_{\text{DNN}} - \beta_0\|_F^2 \xrightarrow{P} 0$.*

The structural model (1.2.1) and the estimator defined in (1.2.5) together yields the estimation error $\{n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top\}^{-1} \{n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i\}$ of $\hat{\beta}_{\text{DNN}}$. The consistency of $\hat{\beta}_{\text{DNN}}$ holds if $n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i \xrightarrow{p} 0$ and $n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top$ is well-defined and invertible as $n \rightarrow \infty$, which are justified by Assumptions 1.2.2 and 1.2.3(b) respectively, under some $g_n^s(Z_i; \hat{W})$ that satisfies Lemma 1.2.2, among other suitable conditions.

Asymptotic normality

Recall that the estimation error of $g_n^s(z; \hat{W})$ for $B_K(z, g)$ and the approximation error of $B_K(z, g)$ for $g(z)$ together constitute the overall error from $g(z)$ to $g_n^s(z; \hat{W})$. While keeping the approximation error of $B_K(z, g)$ dominated with a sufficiently large \bar{K} , we construct the asymptotic normality using the leading term from the estimation error of $g_n^s(z; \hat{W})$ under the high-dimensional central limit theorem (CLT) by Chernozhukov et al. (2017), which relies on the polynomial format of $g_n^s(Z_i; \hat{W})$.

Technically, the Bernstein polynomial approximation applies to any continuous function; however, the feasibility of achieving a good approximation can be significantly affected by the level of smoothness of the underlying function g . For example, to achieve $\varepsilon = o_p(n^{-1/4})$, \bar{K} needs to be at least of order $\delta^{-2} n^{1/4}$ by Lemma 1.2.1. Such \bar{K} can vary dramatically with the level of δ required under the given ε , which may result in unrealistically large orders of the approximating polynomials. One way to avoid such high order polynomials while still achieve the desired convergence rate is to only consider the underlying functions g that are sufficiently smooth. Due to the property of $B_K(Z_i, g_r)$, $\varepsilon = o_p(\bar{K}^{-1})$ holds if g_r is twice differentiable. Then $\bar{K} \sim n^{1/4}$ suffices $\varepsilon = o_p(n^{-1/4})$.

In the current chapter, we keep the smoothness condition for g and approach the asymptotic normality through the high-dimensional CLT by Chernozhukov et al. (2017). Then the following assumptions are required.

Assumption 1.2.4. *Let deg_{poly} be the degree of the polynomial activation function. We have*

$$\frac{\left[d_Z + (\text{deg}_{\text{poly}})^L \right]!}{d_Z! \left[(\text{deg}_{\text{poly}})^L \right]!} = \mathcal{O}(\exp(Cn^C)), \text{ for some constant } C > 0.$$

As demonstrated in Chernozhukov et al. (2017), a Gaussian approximation can be achieved for the sampling distribution of independent random vectors with dimensions increasing no faster than $\exp(Cn^C)$ as the sample size n grows. The architecture of the DNNs is regulated by this constraint for the derivation of the asymptotic normality, whence the number of terms in the polynomial $g_n^s(Z_i; \hat{W})$ should be bounded by $\exp(Cn^C)$. Since the polynomial format of $g_n^s(Z_i; \hat{W})$ under the given structure with a degree- deg_{poly} polynomial activation function involves d_Z variables and $(\text{deg}_{\text{poly}})^L$ degrees overall, the number of terms of such a polynomial can be calculated by the binomial coefficient $\binom{d_Z + (\text{deg}_{\text{poly}})^L}{(\text{deg}_{\text{poly}})^L}$, which is $\left\{ \left[d_Z + (\text{deg}_{\text{poly}})^L \right]! \right\} / \left\{ d_Z! \left[(\text{deg}_{\text{poly}})^L! \right] \right\}$ terms in total, the constraint in Assumption 1.2.4 is required.

Moreover, an additional assumption on higher cross moments between the IVs and the error term U_i is stated in Assumption 1.2.5 to ensure a finite variance of the asymptotic distribution for $\hat{\beta}_{\text{DNN}}$.

Assumption 1.2.5. For $i = 1, \dots, n$, we have $\mathbb{E}[(\prod_{j=1}^{d_Z} Z_i^{\zeta_j})^2 U_i^2] < \infty$ for all $\sum_{j=1}^{d_Z} \zeta_j \leq (\text{deg}_{\text{poly}})^L$.

Then we have the following theorem for the asymptotic normality.

Theorem 1.2.2. Suppose (1.2.1) is correctly specified, and Assumptions 1.2.1 to 1.2.5 hold. With \bar{K} satisfying Assumption 1.2.3(b), $g_n^s(Z_i; \hat{W})$ satisfying Lemma 1.2.2, and $\hat{\beta}_{\text{DNN}}$ as in (1.2.5),

$$\sqrt{n} \left(\Sigma_{BX,n}^{-1} \Sigma_{B,n} \Sigma_{BX,n}^{-\top} \right)^{-1/2} \left(\hat{\beta}_{\text{DNN}} - \beta_0 \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma_{UU} \mathbf{I}_{d_Z} \right),$$

where $\Sigma_{BX,n} := n^{-1} \sum_{i=1}^n [B_K(Z_i, g) X_i^\top]$ and $\Sigma_{B,n} := n^{-1} \sum_{i=1}^n [B_K(Z_i, g) B_K(Z_i, g)^\top]$.

Under the i.i.d. setting, we have that $\Sigma_{BX,n} \xrightarrow{p} \Sigma_{BX}$ and $\Sigma_{B,n} \xrightarrow{p} \Sigma_B$. Hence, the inflator of the estimation error $\hat{\beta}_{\text{DNN}} - \beta_0$ is of order \sqrt{n} , as shown in Theorem 2, which further indicates that $\hat{\beta}_{\text{DNN}}$ is \sqrt{n} -consistent under proper conditions. Also, as noted above, the asymptotics are achieved through an estimation of the polynomial representation $B_K(Z_i, g)$ by the DNN estimator $g_n^s(Z_i; \hat{W})$ and an approximation to the true underlying function g by the polynomial representation $B_K(Z_i, g)$. Theorem 2 states the asymptotic

normality of the leading error of $g_n^s(Z_i; \hat{W})$, while the approximation error ϵ of $B_K(Z_i, g)$ can be made arbitrarily small by selecting a sufficiently large \bar{K} .

1.2.3 Generalized parametric structure model

The proposed estimator is motivated under linear structural models as in (1.2.1), while such DNN-based estimation can adapt to more general parametric models through the generalized method of moments. Essentially, one will fit the Jacobian of the model identification that involves IVs, and such a Jacobian needs to satisfy certain rank conditions to ensure the relevance of the IVs, which further contributes to a sufficiently strong model identification.

Consider any parametric functional form $f(X, \beta_0)$, as such

$$Y = f(X, \beta_0) + U. \quad (1.2.1^*)$$

Note that the function g in the first stage (1.2.2) can be interpreted as a process of obtaining d_X optimized IVs for a just-identified IV regression. However, in a general parametric model, the number of explanatory variables d_X is not necessary equal to the number of parameters, say d_R . Therefore, here we define a function $\gamma : [0, 1]^{d_Z} \rightarrow \mathbb{R}^{d_R}$ of the same concept as g , with a corresponding DNN fitting $\gamma_n^s(Z_i; \hat{W})$ and Bernstein polynomial approximations $B_K(Z_i, \gamma)$, whence the DNN-based estimator, denoted by $\tilde{\beta}_{\text{DNN}}$, can be obtained by matching the population moment $\mathbb{E}[\gamma(Z)[Y - f(X, \beta_0)]] = 0$ with its sample analog, such that

$$\frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \tilde{\beta}_{\text{DNN}})\} = 0. \quad (1.2.6)$$

Then the consistency of such $\tilde{\beta}_{\text{DNN}}$ holds under the following assumptions.

Assumption 1.2.6. *Let B be a compact subspace in \mathbb{R}^{d_R} . There exists $\beta_0 \in B$, such that*

(a). $f(X_i, \beta)$ is measurable in $X_i \in \mathbb{R}^{d_X}$ and continuous in $\beta \in B$.

(b). $\mathbb{E}[\gamma(Z_i)\{Y_i - f(X_i, \beta_0)\}] = 0$, yet $\mathbb{E}[\gamma(Z_i)\{Y_i - f(X_i, \beta)\}] \neq 0$ for all $\beta \neq \beta_0$.

(c). $\mathbb{E}[\sup_{\beta \in B} \|\gamma(Z_i)\{Y_i - f(X_i, \beta)\}\|_F] < \infty$.

Assumption 1.2.6 states the identification of the model under the general parametric functional form $f(X_i, \beta)$, as well as other constraints that indicate the existence and uniqueness of the estimator. Specifically, part (a) ensures that $n^{-1} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W})\{Y_i - f(X_i, \beta)\}$ is well-defined and continuous in β , where the continuity condition is to simplify the derivation of the estimator and its properties. Part (b) implies the uniqueness of the solution to (1.2.6). Part (c) states the moment constraint for the convergence of the sample counterpart. Then the consistency of $\tilde{\beta}_{\text{DNN}}$ can be established as follow.

Theorem 1.2.3. *Suppose Assumptions 1.2.1, 1.2.2, 1.2.3(a), and 1.2.6 hold. With some sufficiently large \bar{K} , and some $\gamma_n^s(Z_i; \hat{W})$ that satisfies Lemma 1.2.2 for γ , there is $\|\tilde{\beta}_{\text{DNN}} - \beta_0\|_F^2 \xrightarrow{p} 0$ for $\tilde{\beta}_{\text{DNN}}$ obtained by solving (1.2.6).*

Recall that for the consistency of $\hat{\beta}_{\text{DNN}}$, Assumption 1.2.2 was imposed for identification, although it was only the moment condition $\mathbb{E}[g(Z_i)U_i] = 0$ that was required. For the consistency of $\tilde{\beta}_{\text{DNN}}$, due to the possible non-linearity in $f(X_i, \beta_0)$, an extra condition that “ $\mathbb{E}[\gamma(Z_i)\{Y_i - f(X_i, \beta)\}] \neq 0$ for all $\beta \neq \beta_0$ ” is included as in Assumption 1.2.6(b), so that the model is well-defined with a unique truth.

Based on the consistency, the asymptotic normality of $\tilde{\beta}_{\text{DNN}}$ can be derived by using the mean-value theorem. First, note that there exists some $\bar{\beta} \in (\min\{\hat{\beta}_{\text{DNN}}, \beta_0\}, \max\{\hat{\beta}_{\text{DNN}}, \beta_0\})$, where the functions “min” and “max” are applied element-wise, such that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \tilde{\beta}_{\text{DNN}})\} = \\ \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \beta_0)\} - \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{\nabla_{\beta} f(X_i, \bar{\beta})\}^{\top} (\tilde{\beta}_{\text{DNN}} - \beta_0). \end{aligned} \quad (1.2.7)$$

Then the asymptotic normality of $\tilde{\beta}_{\text{DNN}}$ follows under the assumptions stated below.

Assumption 1.2.7. *Let B be a compact subspace in \mathbb{R}^{d_R} . There exists $\beta_0 \in B$, such that*

(a). β_0 is in the interior of B .

(b). $f(X_i, \beta)$ is continuously differentiable in $\beta \in B$.

- (c). $n^{-1} \sum_{i=1}^n \gamma(Z_i) \{\nabla_{\beta} f(X_i, b)\}^{\top} \xrightarrow{p} \mathbb{E}[\gamma(Z_i) \{\nabla_{\beta} f(X_i, \beta_0)\}^{\top}]$ for any consistent estimator b of β_0 , where $\mathbb{E}[\gamma(Z_i) \{\nabla_{\beta} f(X_i, \beta_0)\}^{\top}]$ and $\mathbb{E}[B_K(Z_i, \gamma) \{\nabla_{\beta} f(X_i, \beta_0)\}^{\top}]$ have full rank.
- (d). $n^{-1/2} \sum_{i=1}^n B_K(Z_i, \gamma) U_i \xrightarrow{d} N(0, \Sigma_{\gamma U})$, for some positive definite matrix $\Sigma_{\gamma U}$.

Among the conditions in Assumption 1.2.7, it is worth noting that part (c) states the full rank condition that corresponds to the one in Assumption 1.2.3(b) to ensure that the IVs are sufficiently relevant and that the model identification is sufficiently strong.

Theorem 1.2.4. *Suppose Assumptions 1.2.1, 1.2.2, 1.2.3(a), 1.2.6, and 1.2.7 hold. With a sufficiently large \bar{K} , some $\gamma_n^s(Z_i; \hat{W})$ satisfying Lemma 1.2.2 for γ , and consistent $\hat{\beta}_{DNN}$ as in Theorem 1.2.3, there is*

$$\sqrt{n} \left(\Sigma_{BX,n}^{-1} \Sigma_{Bu,n} \Sigma_{BX,n}^{-\top} \right)^{-1/2} \left(\hat{\beta}_{DNN} - \beta_0 \right) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_{d_Z}),$$

where we (re)define the followings $\Sigma_{BX,n} := n^{-1} \sum_{i=1}^n [B_K(Z_i, \gamma) \{\nabla_{\beta} f(X_i, \beta_0)\}^{\top}]$ and $\Sigma_{Bu,n} := n^{-1} \sum_{i=1}^n [B_K(Z_i, \gamma) U_i^2 B_K(Z_i, \gamma)^{\top}]$.

The asymptotic distribution in Theorem 1.2.4 is in a more general form compared with the one in Theorem 1.2.2. Specifically, $\Sigma_{BX,n}$ here will reduce to $n^{-1} \sum_{i=1}^n [B_K(Z_i, g) X_i^{\top}]$ in Theorem 1.2.2 when $f(X, \beta)$ is linear in β , and $(\Sigma_{BX,n}^{-1} \Sigma_{Bu,n} \Sigma_{BX,n}^{-\top})$ here will reduce to $\sigma_{UU} \Sigma_{BX}^{-1} \Sigma_B \Sigma_{BX}^{-\top}$ from Theorem 1.2.2 under homoskedasticity.

The above discussion regarding the generalized parametric models is to demonstrate the adaptability of the proposed method to a more general setting as needed. Despite the results provided in Theorems 1.2.3 and 1.2.4, the current chapter will focus on linear structural models.

1.3 Tests for Weak Identification

The estimator proposed above is intended for cases where the IVs are strong enough under the true first-stage relationship g , but the established asymptotics will collapse if the IVs are not strong enough in any format. Hence, we develop a hypothesis test based on the

first-stage DNN fitting for weak identification and determine whether the IVs are jointly strong enough to provide valid causal inference. This test is motivated under the linear structural model as in (1.2.1). One can also extend the test for a generalized parametric structural model under some proper local linearization; however, we will omit extensive discussion from here as it is not the focus of the current chapter.

The concentration parameter is an established measurement in literature for the instruments' overall explanatory power under the given first-stage model specification and applied to determine how strong the model is identified (e.g., Staiger and Stock, 1997; Stock et al., 2002; Hansen et al., 2008; Andrews et al., 2019). In line with the concept of the concentration parameter, we consider a measurement of concentration μ_l^2 for X_l and μ_{nl}^2 for X_{nl} , such that

$$\mu_l^2 := \frac{\text{tr}(\pi_0^\top \mathbb{E}[Z_i Z_i^\top] \pi_0)}{\text{tr}(\text{Var}(V_i))}, \quad \mu_{nl}^2 := \frac{\text{tr}(\text{Var}(X_{nl}(Z_i)))}{\text{tr}(\text{Var}(V_i))}.$$

As shown in previous studies (e.g., Chao and Swanson, 2005), the conventional 2SLS estimator with a linear first stage has overlooked X_{nl} , and the consistency of this estimator requires $d_Z/(n\mu_l^2) \rightarrow 0$. Such a convergence condition delivers three messages. First, with a fixed number of instruments d_Z , a diverging $n\mu_l^2$ as $n \rightarrow \infty$ implies that the explanatory power of the instruments via X_l remains strong with an increasing sample size, while a bounded $n\mu_l^2$ as $n \rightarrow \infty$ implies that the instruments are asymptotically irrelevant, whence the model is weakly identified. Second, even a model has $n\mu_l^2 \rightarrow \infty$, it can only tolerate so fast of a growth in the number of instruments before it becomes weakly identified. Third, when the true relationships between X and Z is hardly linear and X_l merely provides a weak identification, a conventional two-stage IV regression with a linear first stage will appear to be weakly identified, yet stronger identification can be achieved through X_{nl} , as long as the explanatory power of Z via X_{nl} stays high enough.

In the spirit of the parameter μ_{nl}^2 , we now construct a test statistic that captures the first-stage signal-to-noise ratio. Since the IVs need to be jointly relevant in each and every entry of $g(Z_i)$, we simplify the discussion to $d_X = 1$ without loss of generality given that d_X is a finite fixed number. Hence, our statistic will be defined for $d_X = 1$, and the test should be applied to every single entry of $g(Z_i)$ — the presence of weak identification for any entry of $g(Z_i)$ indicates that better IVs are needed. We now define our statistic GF ,

for a generalized first-stage F-test, as follow:

$$GF = \frac{\sum_{i=1}^n \{g_n^s(Z_i; \hat{W})X_i\} / p_L}{n^{-1} \sum_{i=1}^n \hat{V}_i^2},$$

where p_L is the width of the last hidden layer in the DNN fitting $g_n^s(Z_i; \hat{W})$, and $\hat{V}_i := X_i - g_n^s(Z_i; \hat{W})$. The behavior of this statistic is addressed in the following theorem.

Theorem 1.3.1. *Suppose the model in (1.2.1) is correctly specified, and Assumptions 1.2.1 to 1.2.5 hold. Then for some finite p_L , under the null hypothesis $H_0 : g = 0$ and the local alternatives of $H_a : g = o_p(n^{-1/2})$, there is $p_L GF \xrightarrow{d} \chi_{p_L}^2$ as $n \rightarrow \infty$; under the local alternatives of $H_a : g \sim n^{-1/2}$, there is $p_L GF \xrightarrow{d} \chi_{p_L}^2(\alpha)$ as $n \rightarrow \infty$, where $\chi_{p_L}^2(\alpha)$ denotes a noncentral chi-squared distribution with the noncentral parameter α and $\alpha \sim n^{-1/2}$; under the alternatives of $H_a : g \notin \mathcal{O}_p(n^{-1/2})$, there is $p_L GF \rightarrow \infty$ as $n \rightarrow \infty$.*

Note that the value p_L takes the place of the first degree of freedom in a traditional F-statistic. More generally, based on the definition of the statistic GF , this degree of freedom should correspond to the dimension of the space of g_n^s . This dimension essentially depends on the construction of the polynomial composition, which determines the dimensions of the projections between layers. However, relying on the degree of freedom from all the projections among all the layers complicates practical application. A simpler approach is to consider only the first $L - 1$ layers to achieve a desired \bar{K} and use a smaller p_L in the last hidden layer to bound the dimension of g_n^s from below. This way, the degree of freedom relies solely on p_L , as stated in the theorem.

In this theorem, other than the null hypothesis of $g = 0$, we also discuss the local alternatives where the IVs are so weakly relevant that the asymptotic distribution of $p_L GF$ still holds chi-squared. While when the IVs get stronger such that the signal g is no longer in a $n^{-1/2}$ neighborhood of 0, $p_L GF$ diverges with n and achieves unit power asymptotically.

However, it is important to note that the rejection, based on the critical value simply obtained from the distribution $\chi_{p_L}^2$ at some given significance level, of the null hypothesis $g = 0$ or the alternatives $g \in \mathcal{O}_p(n^{-1/2})$ only indicates $g \notin \mathcal{O}_p(n^{-1/2})$ but not necessarily that the IVs are strong enough. Despite widely applying the first-stage F-test for IV relevance, literature has pointed out that a traditional F-test on the null of $g = 0$ is inadequate with

weak IVs, since the low relevance of IVs leads to tests with large size distortions (e.g., Hall et al., 1996; Stock et al., 2002; Stock and Yogo, 2002; Lee et al., 2021). Some studies have proposed to adjust the test criteria to allow for valid inference. For example, Stock and Yogo (2002) tabulate critical values to enable the first-stage F-test for weak IVs. While Lee et al. (2021) argue that practitioners have been erroneously referring to Stock and Yogo (2002) tables and using the rule-of-thumb threshold of 10 to construct confidence intervals for the IV estimators by mistake, whereas truthfully, a usual 95% confidence interval for the IV estimator requires a threshold of 104.7 instead of 10 for the first-stage F-statistic. Some other studies have also suggested to test the null hypotheses such as “ $n\mu_{nl}/d_Z$ is bounded above by some weak IV threshold” against the alternatives that “ $n\mu_{nl}/d_Z$ exceeds the threshold” (e.g., Stock et al., 2002; Sanderson and Windmeijer, 2016), where the threshold is determined by any given level of IV estimation bias.

In our case, we develop the test for weak IVs, following the suggested adjustment by Stock and Yogo (2002) (Lee et al., 2021) and adopting a criterion corresponding to their corrected threshold of 104.7. To determine our test criterion, first note that for an F -distribution with the degree of freedom df , $dfF \xrightarrow{d} \chi_{df}^2$ as $n \rightarrow \infty$. Taking the threshold of 104.7 as an example, the corresponding criterion in the limiting chi-squared distribution will be $104.7df$. Hence, we will reject the hypothesis of weak identification if $p_L GF > 104.7p_L$, and fail to reject otherwise. If the weak identification hypothesis is rejected, then we suggest to use our method proposed in Section 1.2 for inference; while if the weak identification hypothesis cannot be rejected, better IVs are needed.

As mentioned above, this test is designed for linear structural models. For a generalized parametric model, the occurrence of weak identification will be the violation of Assumption 1.2.6(a), which can be tested through the full rank condition in Assumption 1.2.7(c). Then the test statistic will be constructed specifically according to the functional form $f(X, \beta)$, which will lead to a different expression from the GF defined above.

1.4 Simulation Analysis

We now use a simulation study to illustrate the proposed DNN-based estimation and hypothesis test.

1.4.1 Global Settings

First, consider the structural model as follows for $i = 1, \dots, n$

$$Y_i = 1 + \beta_0 X_i + U_i,$$

with

$$\begin{aligned} \beta_0 &= 3, \\ U_i &\sim \Gamma(2, \sqrt{10}) - 2\sqrt{10}. \end{aligned}$$

We include one endogenous explanatory variable X_i and five exogenous IVs denoted by the 5-vector Z_i , whence the first stage is as such

$$X_i = \frac{1}{\sqrt{n}} \sum_{d=1}^5 Z_i^{(d)} + \frac{2}{5} \sum_{d=1}^5 \left(Z_i^{(d)}\right)^2 + \frac{2}{15} \sum_{d=1}^5 \left(Z_i^{(d)}\right)^4 + \frac{2}{75} \sum_{d=1}^5 \left(Z_i^{(d)}\right)^6 + V_i,$$

with

$$Z_i \sim \mathcal{N}(\mathbf{0}, I_5), \quad V_i \sim \Gamma(2, \sqrt{5}) - 2\sqrt{5},$$

where I_5 denotes a 5-by-5 identity matrix, $Z_i^{(d)}$ denotes the d th entry of Z_i , and the endogeneity of X_i is imposed by letting $\text{corr}(U_i, V_i) = 0.8$.

1.4.2 Simulation Designs

In **Design (i)**, a sample of size $n = 2000$ is generated repeatedly for $S = 1000$ times based on the above data generating process (DGP). The first stage consists of a linear component $n^{-1/2} \sum_{d=1}^5 Z_i^{(d)}$ of the order $n^{-1/2}$ and a non-linear part of $\mathcal{O}_p(1)$. To observe the issue caused by ignoring the non-linearity and demonstrate how the DNN-based estimator provides a solution, we let the non-linear components take the second, fourth and sixth degrees of monomials of the IVs, so that a linear specification cannot pick up sufficient information from the non-linear components due to orthogonality, and thus, remains weak.

As shown in literature, the traditional 2SLS estimator will be biased towards the OLS

estimator if the IVs are weak under the parametric specification (e.g., Stock and Yogo, 2002; Sanderson and Windmeijer, 2016). In this design, we first compare the performance of the proposed DNN-based estimator with the traditional 2SLS in terms of their biases relative to the OLS estimator to demonstrate that the proposed estimator can help correct this bias by enhancing the IVs through potential non-linear relationships in the first stage.

Under the weak instrument conditions, there are several alternative estimators, which have a smaller bias and better small sample properties than 2SLS. First we choose the limited information maximum likelihood (LIML) estimator. Forchini and Jiang (2019) stated that, compared with 2SLS, LIML is a linear combination of the OLS and 2SLS estimators and consider the weights biased on the data. Such weights could eliminate the 2SLS bias with weak instruments. Apart from LIML, we apply a Sieve estimation in the first stage and obtain a Sieve-based estimator in the linear structure model. The Sieve-based estimator could capture more information of weak instruments in the first stage with a non-parametric model. Also, the sieve estimation induce well-posed optimization problems and allow consistent estimation of both parametric and non-parametric parts when becoming sufficiently rich and dense in the whole space with an enlarging sample (e.g., Chen, 2007; Chen and Liao, 2015). In this design, we then compare the performance of the proposed DNN-based estimator with the LIML estimator and Sieve-based estimator in terms of their biases relative to the OLS estimator to demonstrate that the proposed estimator can obtain a relatively small bias and a relatively small standard error than the traditional estimators under the weak instrument conditions.

In **Design (ii)**, a relatively small sample of size $n = 1000$ is generated repeatedly for $S = 1000$ times based on the above data generating process (DGP). The first stage also consists of a linear component $n^{-1/2} \sum_{d=1}^5 Z_i^{(d)}$ of the order $n^{-1/2}$ and a non-linear part of $\mathcal{O}_p(1)$. We still let the non-linear components take the second, fourth and sixth degrees of monomials of the IVs. In this design, we compare the performance of the proposed DNN-based estimator with the traditional 2SLS in terms of their biases relative to the OLS estimator, to demonstrate that the proposed estimator can also help correct this bias in a relatively small sample size. The robustness of our proposed estimator is also examined in Design (ii).

In **Design (iii)**, different neural network structures are discussed and compared. The upper bound between estimated function and underlying function is correlated with the

degree of polynomial approximation used in activation function, and the number of hidden layers. The optimal approximation power is reached only if the neural network have optimal architectures (Bartan and Pilanci, 2021). In this design, we use the same data sample in Design (i) and compare the performance of the proposed DNN-based estimator under different neural network architectures: one hidden layer with low-degree polynomial approximation activation function (labeled as DNN-L1 estimator); two hidden layer with low-degree polynomial approximation activation function (labeled as DNN-L2 estimator); one hidden layer with high-degree polynomial approximation activation function (labeled as DNN-H1 estimator); two hidden layer with high-degree polynomial approximation activation function (labeled as DNN-H2 estimator). Note that our proposed DNN estimator without any additional assumption is a DNN-H2 estimator.

In **Design (iv)**, the performance of our proposed DNN estimator under strong instrument conditions is examined with 2SLS and OLS estimator. We replace the first stage in the global settings by,

$$X_i = \sum_{d=1}^5 Z_i^{(d)} + \frac{2}{5} \sum_{d=1}^5 \left(Z_i^{(d)} \right)^2 + \frac{2}{15} \sum_{d=1}^5 \left(Z_i^{(d)} \right)^4 + \frac{2}{75} \sum_{d=1}^5 \left(Z_i^{(d)} \right)^6 + V_i.$$

A sample of size $n = 2000$ is generated repeatedly for $S = 1000$ times based on the above data generating process (DGP). The first stage consists of a strong linear component $\sum_{d=1}^5 Z_i^{(d)}$, under which the instruments are strong as there is a obvious linear relationship between the endogenous variable and the instruments. We also let the non-linear components take the second, fourth and sixth degrees of monomials of the IVs. We compare the performance of the proposed DNN- based estimator with the traditional 2SLS in terms of their biases relative to the OLS estimator to demonstrate that the proposed estimator is a more efficient estimator under strong instrument conditions.

1.4.3 Estimation

In practice, we need to determine the activation functions and the DNN structures before we proceed with the estimation. To define the polynomial activation functions, note that there are various activation functions with appealing advantages in terms of their specifications (or shapes), and we are particularly interested in the polynomial activation functions

due to their properties as polynomials. Therefore, if there are any certain activation functions that have good performance in general, we can find a polynomial approximation for those activation functions and then use this polynomial approximation as the polynomial activation function for training. By doing so, the general shape of that activation function of interest can be preserved, and the properties of polynomials can also be imposed. In this simulation analysis, we use a poly-Sigmoid activation function, which is a degree-seven polynomial approximation of the Sigmoid activation function.

As for the DNN structure, in this simulation study, we employ a network with two hidden layers, where the first hidden layer includes 150 units and the second includes 75. To obtain a more thorough search for the optimal DNN structure in practice, one can make it data-driven from a broad pool of candidates as needed by the complexity of the problems. To train the DNN under any given structure, we split the total sample of n into 80% of training sample and 20% of validation sample, where the training sample is to obtain a $g_n^s(Z_i; \hat{W})$ by minimizing the loss function, and the validation sample is to monitor the out-of-sample performance for possible over-fitting. After a sufficiently large number of iterations, the trained model from the iteration with the smallest validation loss will be selected for the first-stage fitting $g_n^s(Z_i; \hat{W})$. Then we can move on to obtain the estimated coefficient $\hat{\beta}_{\text{DNN}}$ by using (1.2.5).

1.4.4 Results and discussion

Design (i)

Among the 1000 rounds of simulation, we obtain the estimated values for the coefficient $\beta_0 = 3$ by using the OLS estimator, the traditional 2SLS estimator, and the proposed DNN-based estimator. For each estimator, we compute the empirical mean, bias, standard error (SE), mean squared error (MSE), and the relative bias to the OLS estimator, as displayed in Table 1.1. The DNN-based estimator is obtained as explained above in Section 1.2, and the traditional 2SLS is based on a linear first-stage, which is only weakly identified under our DGP. The OLS is for the estimation of the structural model without accommodating the endogeneity.

As shown in Table 1.1, the relative bias of the DNN-based estimator is around 6.18%, which is much smaller than the relative bias of the 2SLS estimator (45.94%). A comparison

Table 1.1: Estimation Statistics for OLS, 2SLS and DNN in Design (i)

| Estimator | Mean | Bias | SE | MSE | Relative bias to OLS | RR(F-test) | RR(GF-test) |
|-----------|--------|--------|--------|--------|----------------------|------------|-------------|
| OLS | 3.2255 | 0.2255 | 0.0533 | 0.0537 | 1 | — | — |
| 2SLS | 3.1036 | 0.1036 | 0.3171 | 0.1113 | 45.94% | 0.0050 | — |
| DNN | 3.0139 | 0.0139 | 0.0240 | 0.0008 | 6.18% | — | 0.9940 |

among their distributions is presented in Figure 1.2(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 1.2(b).

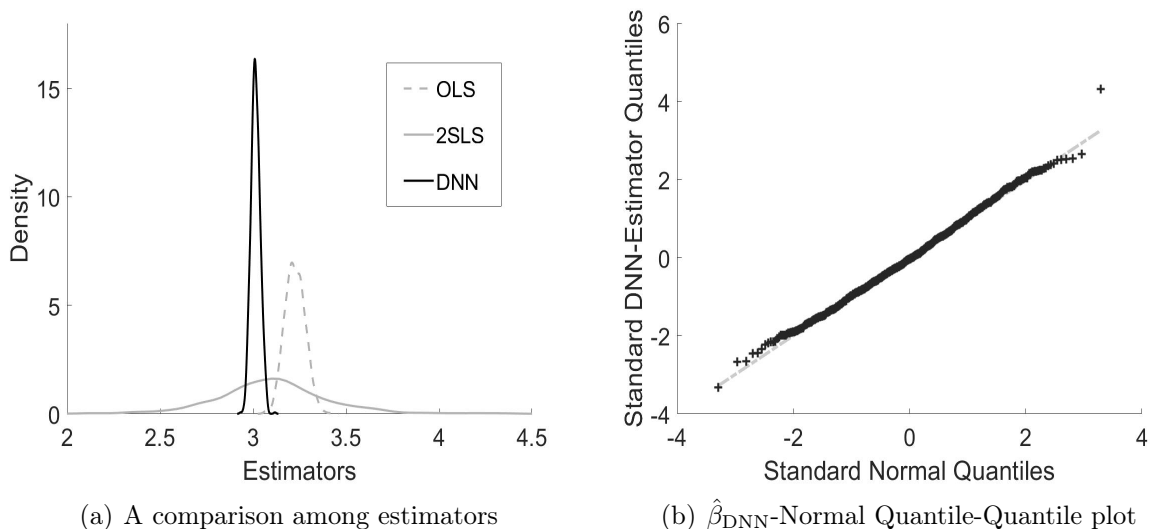


Figure 1.2: Distributions of the OLS, 2SLS and DNN in Design (i)

Furthermore, we also report the results of the traditional first-stage F-test for the 2SLS estimator and our generalized first-stage F-test for the DNN-based estimator. Since the DGP is designed so that the IVs are weak under a linear first stage but are not weak under the true relationship $g(Z)$, the traditional first-stage F-test for 2SLS should fail to reject the null of weak IVs, and our generalized F-test for DNN should have a rejection rate converging to unity. Hence, The results in the last column of Table 1.3 – the rejection rate of the corresponding first-stage F-test and general F-test (if applicable) – are as expected. For 2SLS, the rejection rate (RR) for F-test is 0.005. For DNN-based estimator, the RR

for generalized F test is 0.994.

Among the 1000 rounds of simulation, we also obtain the estimated values for the coefficient $\beta_0 = 3$ by using the LIML estimator and the Sieve-based estimator. For the Sieve-based estimator, we run a sieve stochastic gradient descent estimator with cosine basis functions in the first stage and obtain the Sieve-based estimator in the linear structure form. For each estimator, we compute the empirical mean, bias, standard error (SE), mean squared error (MSE), and the relative bias to the OLS estimator, as displayed in Table 1.2. The DNN-based estimator is obtained as explained above in Section 1.2, and the

Table 1.2: Estimation Statistics for OLS, LIML, Sieve and DNN in Design (i)

| Estimator | Mean | Bias | SE | MSE | Relative bias to OLS | RR(GF-test) |
|-------------|--------|--------|--------|--------|----------------------|-------------|
| OLS | 3.2255 | 0.2255 | 0.0533 | 0.0537 | 1 | — |
| LIML | 3.0436 | 0.0436 | 0.5115 | 0.2635 | 19.35% | — |
| Sieve-based | 3.0260 | 0.0260 | 0.1036 | 0.0030 | 11.56% | 0.9760 |
| DNN | 3.0139 | 0.0139 | 0.0240 | 0.0008 | 6.18% | 0.9940 |

traditional 2SLS is based on a linear first-stage, which is only weakly identified under our DGP. The OLS is for the estimation of the structural model without accommodating the endogeneity.

As shown in Table 1.2, the relative bias of the DNN-based estimator is around 6.18%, which is much smaller than the relative bias of the LIML estimator (19.35%) and the Sieve-based estimator (11.56%). A comparison among their distributions is presented in Figure 1.3. The evidence in Table 1.2 illustrates that DNN have a larger flexibility and potentially a faster convergence rate than traditional estimators of LIML and Sieve. As we explained before, DNN offer a “multi-layer” extension of the traditional sieve regressions, and they are capable of improving on a larger scale with a growing sample size compared to traditional single-layer learning methods (e.g., Fabozzi et al., 2019; Shen et al., 2023; Husmeier, 2012). Given a better estimation in the first stage by DNN estimation, our proposed estimator is more efficient than other two traditional estimators. As the non-linear components in the first stage take the second, fourth and sixth degrees of monomials of the IVs, the bias between DNN estimator and other classic estimators is not significant. Given a more complex non-linear function in the first stage, the DNN estimator could perform much better than other classic estimators.

Also, we report the results of our generalized first-stage F-test for the Sieve-based estimator and DNN-based-estimator. Here, we consider the p_L as the number of basic functions for Sieve-based estimator. The results in the last column of Table 1.3 – the rejection rate of the corresponding first-stage F-test and general F-test (if applicable) – are as expected. For Sieve-based estimator, the RR for generalized F test is 0.976. For DNN-based estimator, the RR for generalized F test is 0.994.

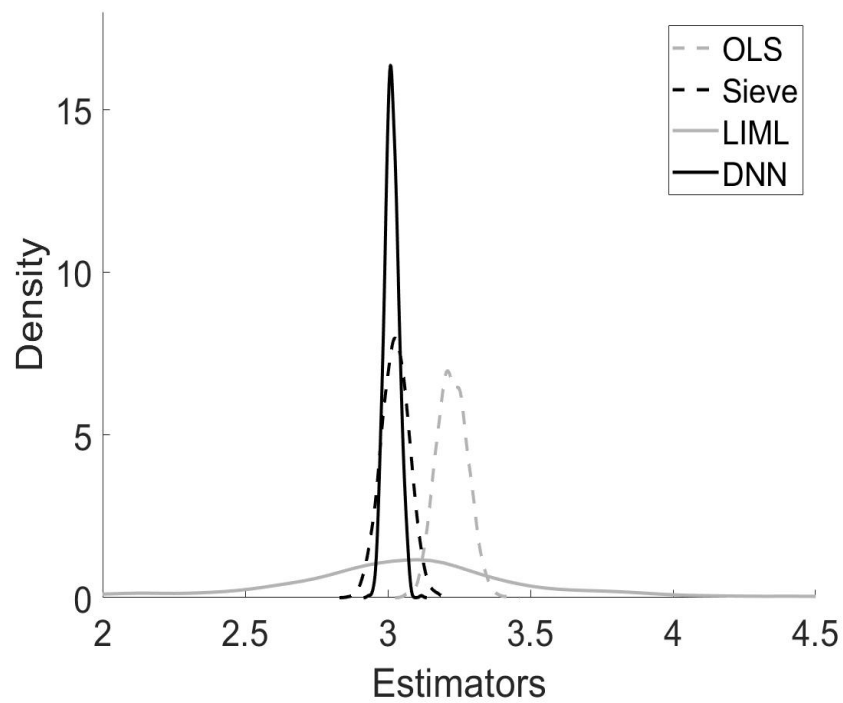


Figure 1.3: Distributions of the OLS, LIML, Sieve and DNN in Design (i)

Design (ii)

With a relatively small sample size $n = 1000$, among the 1000 rounds of simulation, we obtain the estimated values for the coefficient $\beta_0 = 3$ by using the OLS estimator, the traditional 2SLS estimator, and the proposed DNN-based estimator. For each estimator, we compute the empirical mean, bias, standard error (SE), mean squared error (MSE), and the relative bias to the OLS estimator, as displayed in Table 1.3.

Table 1.3: Estimation Statistics for OLS, 2SLS and DNN in Design (ii)

| Estimator | Mean | Bias | SE | MSE | Relative bias to OLS | RR(F-test) | RR(GF-test) |
|-----------|--------|--------|--------|--------|----------------------|------------|-------------|
| OLS | 3.2321 | 0.2321 | 0.0711 | 0.0558 | 1 | — | — |
| TOLS | 3.1024 | 0.1024 | 0.3082 | 0.1055 | 44.13% | 0.0070 | — |
| LIML | 3.0436 | 0.0436 | 0.5115 | 0.2635 | 19.35% | — | — |
| Sieve-SGD | 3.0263 | 0.0263 | 0.1168 | 0.0014 | 11.33% | — | 0.9630 |
| DNN | 3.0211 | 0.0211 | 0.0264 | 0.0011 | 9.10% | — | 0.9920 |

As shown in Table 1.3, the relative bias of the DNN-based estimator is around 9.10%, which is much smaller than the relative bias of the 2SLS estimator (44.13%). A comparison among their distributions is presented in Figure 1.4(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 1.4(b).

We also report the results of the traditional first-stage F-test for the 2SLS estimator and our generalized first-stage F-test for the DNN-based estimator. The results in the last column of Table 1.3 – the rejection rate of the corresponding first-stage F-test and general F-test (if applicable) – are as expected. For 2SLS, the rejection rate (RR) for F-test is 0.007. For DNN-based estimator, the RR for generalized F test is 0.994.

Design (iii)

In this section, we compare the performance of the proposed DNN-based estimator under different neural network architectures. For DNN-L1 and DNN-H1 estimator, the hidden layer includes 150 units. For DNN-L2 and DNN-H2, the first hidden layer includes 150 units and the second includes 75. Note that we use the same data sample in Design (i) and the same results for DNN-H2 estimator. Among the 1000 rounds of simulation, we obtain the estimated values for the coefficient $\beta_0 = 3$ by using the OLS estimator, DNN-L1,

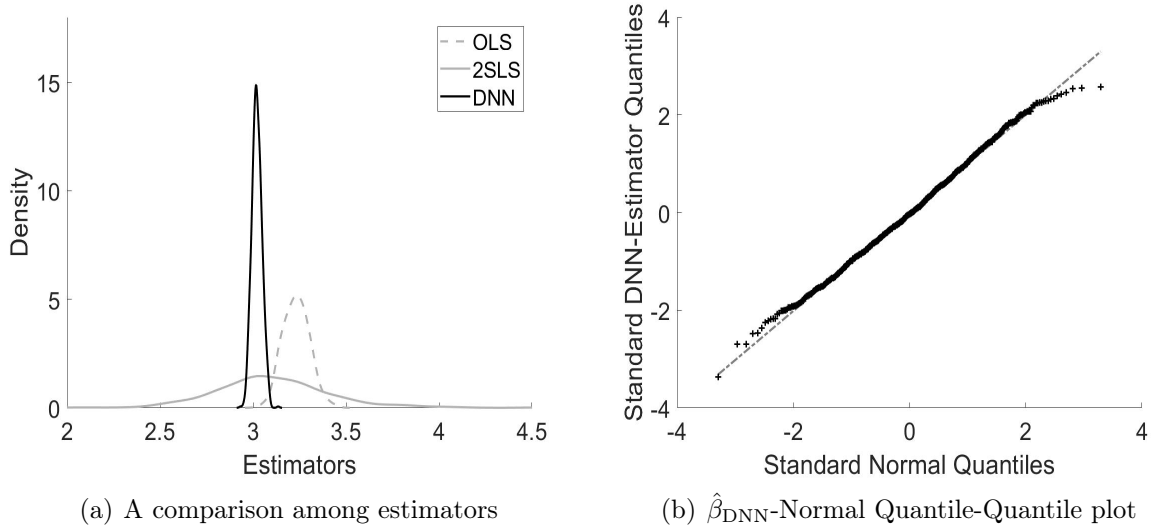


Figure 1.4: Distributions of the OLS, 2SLS and DNN Estimators in Design (ii)

DNN-L2 and DNN-H1 estimator, along with the results of DNN-H2 estimator in Table 1.1. For each estimator, we compute the empirical mean, bias, standard error (SE), mean squared error (MSE), and the relative bias to the OLS estimator, as displayed in Table 1.4.

Table 1.4: Estimation Statistics for OLS and DNN estimators in Design (iii)

| Estimator | Mean | Bias | SE | MSE | Relative bias to OLS | RR(GF-test) |
|-----------|--------|--------|--------|--------|----------------------|-------------|
| OLS | 3.2255 | 0.2255 | 0.0533 | 0.0537 | 1 | — |
| DNN-L1 | 3.0242 | 0.0242 | 0.0329 | 0.0017 | 10.73% | 0.9830 |
| DNN-L2 | 3.0203 | 0.0203 | 0.0284 | 0.0012 | 8.99% | 0.9890 |
| DNN-H1 | 3.0142 | 0.0142 | 0.0251 | 0.0008 | 6.32% | 0.9920 |
| DNN-H2 | 3.0139 | 0.0139 | 0.0240 | 0.0008 | 6.18% | 0.9940 |

As shown in Table 1.4, the relative bias of the DNN-H2 estimator is around 6.18%, which is smaller than the relative bias of the DNN-L1 estimator (10.73%), the DNN-L2 estimator (8.99%), and the DNN-H1 estimator (6.32%). A comparison among their distributions is presented in Figure 1.5(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in

Figure 1.5(b), 1.5(c) and 1.5(d). We also report the results of our generalized first-stage F-test for the DNN-based estimators. The rejection rate (RR) of GF-test for DNN-L1, DNN-L2, DNN-H1 and DNN-H2 are 0.983,0.989,0.992,and 0.994.

We notice that DNN-H1 and DNN-H2 estimator are more efficient than the rest of estimators, and the performances of DNN-H1 and DNN-H2 are very close. We conclude that (1) DNN structures with high-degree polynomial approximation function have a better performance among the DGP in Design (iii); (2) DNN structures with more hidden layers have a better performance of the DGP in Design (iii) when the activation function is a low-degree polynomial approximation; (3) DNN structures with additional hidden layer does not significantly improve the efficiency of the estimator with high-degree polynomial approximation activation, as the complexity of the DNN structure is sufficient to the DGP in Design (iii) and the default structure of DNN (two hidden layers with high-degree polynomial approximation activation) is approximately the optimal DNN architecture for the estimation. These conclusions illustrate the robustness of our proposed DNN estimator over different network architectures, and further verify the reliability of the results we obtained in Design (i) and (ii).

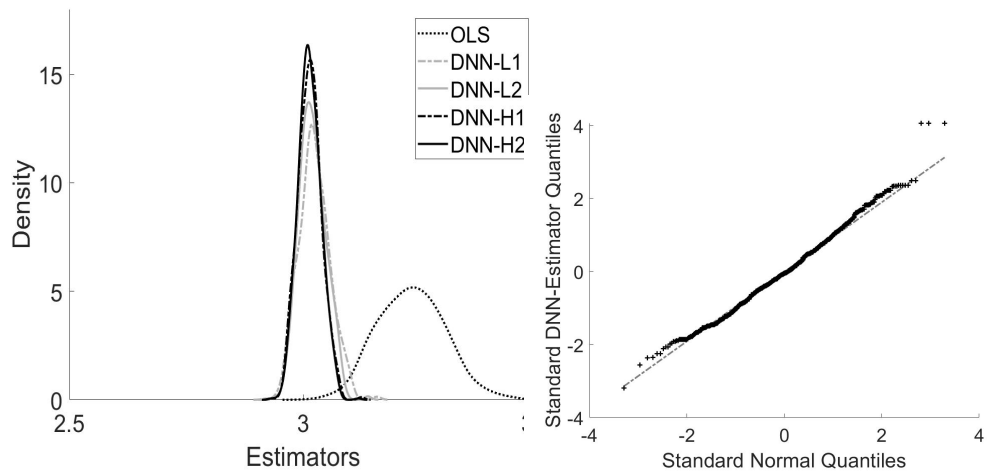
Design (iv)

The aim of Design (iv) is to compare the performance of the proposed DNN-based estimator with the traditional 2SLS in terms of their biases relative to the OLS estimator to demonstrate that the proposed estimator is a more efficient estimator under strong instrument conditions. Among the 1000 rounds of simulation, we obtain the estimated values for the coefficient $\beta_0 = 3$ by using the OLS, 2SLS and proposed DNN estimator. For each estimator, we compute the empirical mean, bias, standard error (SE), mean squared error (MSE), and the relative bias to the OLS estimator, as displayed in Table 1.5.

Table 1.5: Estimation Statistics for OLS, 2SLS, and DNN in Design (iv)

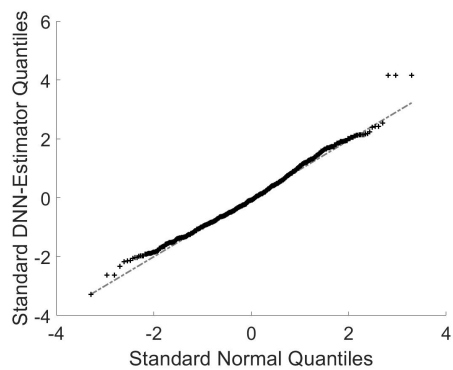
| Estimator | Mean | Bias | SE | MSE | Relative bias to OLS | RR(F-test) | RR(GF-test) |
|-----------|--------|--------|--------|--------|----------------------|------------|-------------|
| OLS | 3.2143 | 0.2143 | 0.0487 | 0.0483 | 1 | — | — |
| 2SLS | 3.0176 | 0.0176 | 0.0908 | 0.0083 | 8.21% | 0.8893 | — |
| DNN-H2 | 3.0113 | 0.0113 | 0.0240 | 0.0008 | 5.27% | — | 0.9970 |

As shown in Table 1.5, the relative bias of our proposed DNN estimator is around 5.27%, which is smaller than the relative bias of the 2SLS estimator (8.21%), while the

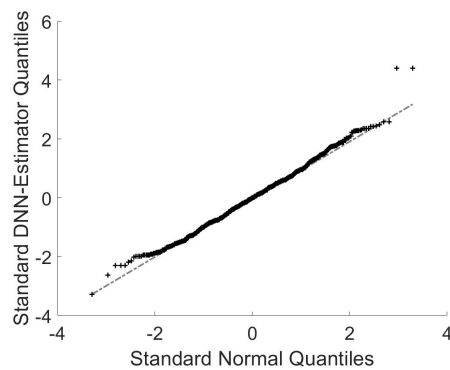


(a) A comparison among estimators

(b) DNN-L1-Normal Quantile-Quantile plot



(c) DNN-L2-Normal Quantile-Quantile plot



(d) DNN-H1-Normal Quantile-Quantile plot

Figure 1.5: Distributions of the DNN Estimators in Design (iii)

standard error of DNN estimator is much smaller than 2SLS estimator. A comparison among their distributions is presented in Figure 1.6(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 1.6(b). We also report the results of traditional F-test for 2SLS estimator and our generalized first-stage F-test for the DNN-based estimators. The rejection rate (RR) of F-test for 2SLS estimator is 0.8893, and the rejection rate (RR) of GF-test for DNN estimator is 0.997.

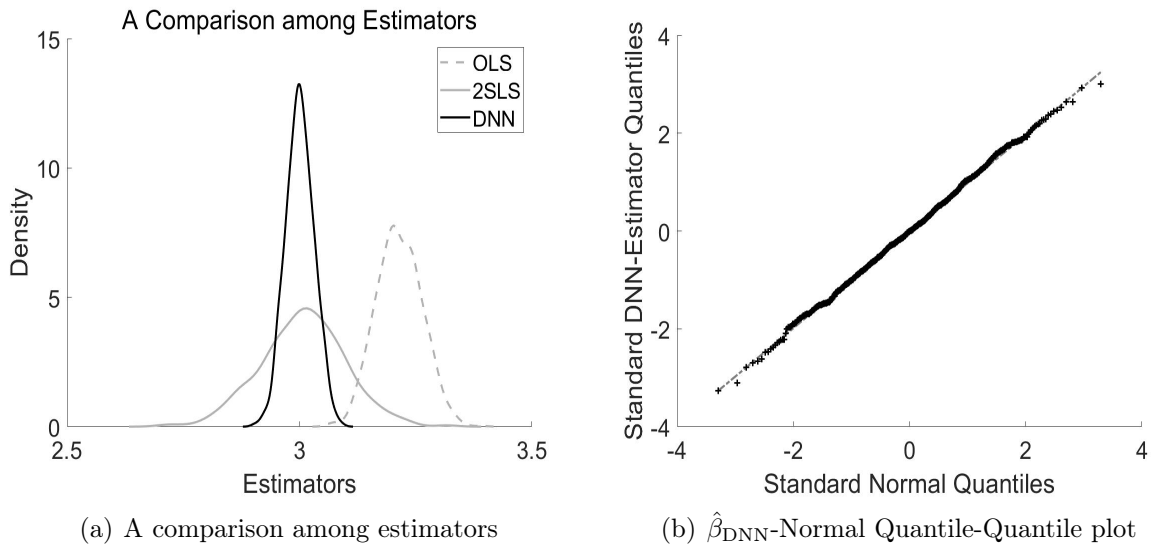


Figure 1.6: Distributions of the OLS, 2SLS and DNN in Design (iv)

We notice that although 2SLS is already a consistent estimator under strong instrument conditions, our proposed DNN estimator is more efficient with a relatively small bias and standard error as more information of the endogenous variable towards the instruments is captured by DNN in the first stage. However, calculating DNN estimator is more complicated and need a relatively long training procedure, while calculating 2SLS is easy and straightforward. This is considered as trade-offs of DNN estimator to approach a better performance under strong instrument conditions.

1.5 Conclusion

The 2SLS regression analysis with IVs are generally applied in causal inference when endogeneity is suspected in the explanatory variables. Traditionally, a parametric functional form will be imposed on the relationship between the IVs and the endogenous explanatory variables. Yet, when such parametric functional form barely captures a weak relationship, which is quite common in practice, the model is only weakly identified, and the subsequent inference of the structural form will not be reliable. If the underlying relationship between the IVs and the endogenous explanatory variables is actually strong enough yet is overlooked by the parametric specification, methods that provide good first-stage fittings without demanding the functional form will be sensible.

In this chapter, we adopt a two-stage estimator by fitting the first stage with a DNN, and we contribute to the literature in two main aspects. First, we derive the large sample properties of a DNN-based estimator formulated on polynomial activation functions, which also helps improve the interpretability of the DNN and the DNN-based estimators. Second, we develop a generalized first-stage F-test for weak IVs and justify the test validity, which provides a guideline on whether we can use a DNN for the first-stage fitting to enhance the IVs' strength, or we simply need better IVs.

Chapter 2

Bias-adjusted Inference with Unobserved Confounding¹

2.1 Introduction

Omitted variable bias constitutes a major concern in causal inference with observational data. Unbiased inference usually relies on the assumption that the treatment is exogenous, conditional on a set of observed control variables. However, the discussion regarding whether the estimated treatment effects are influenced by unobserved confounding is often fundamentally unverifiable in practice (e.g., Rosenbaum, Paul R and Rubin, Donald B, 1983; Pearl, 1995; Imbens and Rubin, 2015).

Such an issue motivates the sensitivity analysis that examines how robust a result is against unobserved confounding (see, e.g., Cornfield et al., 1959; Rosenbaum and Rubin, 1983) and the bias-adjusted inference that, to some degree, accommodates the confounding effects (see, e.g., Riegg, 2008; Oster, 2019; Jesson et al., 2021; Wüthrich and Zhu, 2023). For example, some studies addressed specific research questions of interest, such as causal risk-ratios and causal risk differences, focusing on developing problem-specific solutions (e.g., Robins, 1999; Frank et al., 2013; Brumback et al., 2004; Altonji et al., 2005; Imai et al., 2010; Dorie et al., 2016; Middleton et al., 2016). Without the problem-specific

¹This chapter is co-authored with Tao Chen and Renfang Tian.

constraints, other recent works discussed the estimation of the bounds of bias subject to confounding effects while making assumptions on the relationship between observed and unobserved confounding (e.g., Frank, 2000; Rosenbaum and Rosenbaum, 2002; Imbens, 2003; Hosman et al., 2010; VanderWeele and Arah, 2011; Frank et al., 2013; Blackwell, 2014; Carnegie et al., 2016; VanderWeele and Ding, 2017; Kallus and Zhou, 2018; Kallus et al., 2019; Kallus, 2023).

To relax the restrictive parametric assumptions, there are also studies proposing reparameterization of the omitted variable bias framework in terms of partial R^2 – accommodating unknown and possibly non-linear functional forms – and showed that the bounds of the omitted variable bias depend only on the confounder’s strength of association with the treatment and the outcome (e.g., Cinelli and Hazlett, 2020; Scharfstein et al., 2021; Bonvini and Kennedy, 2022). Another approach suggests the use of the non-parametric R^2 , introduced by Pearson (1905) as a generalization of the linear R^2 , which has been studied in non-parametric regression (see, e.g., Doksum and Samarov, 1995) and applied in place of the linear R^2 to generalize the bounds of omitted variable bias for partially linear models (Chernozhukov et al., 2022).

Apart from estimating the bounds of the bias, there has also been literature on developing the bias reduction or bias correction methods. For example, the proportion of variation in the outcome, or in the conditional expectation of the outcome explained by the unobserved confounding in a non-parametric model, could contribute to estimate strength of associated unobserved confounding factors over the treatment effect, and obtain a more efficient inference for the estimator under restrictive conditions (e.g., Wilms et al., 2021; Chernozhukov et al., 2022). Wüthrich and Zhu (2023) proposed a Lasso-based inference methods to modify the high dimensional OLS-based inference and reduced the bias over unobserved confounding factors. Kernel regression and kernel regularized least squares (KRLS) have been studied and applied by Hainmueller and Hazlett (2014), to reduce the mis-specification bias over unobserved variables, avoiding strong parametric assumptions. Machine learning methods, such as deep neural networks (DNNs), were also competitive candidates for reducing the bias over noises and estimating the the conditional expectation function of outcome over interested variables (Dangeti, 2017).

In line with the estimation of classical non-parametric regression models that may involve methods such as spline regression, local polynomial regression, and kernel regression,

etc, a sieve estimator can be constructed as a collection of subsets of finite-dimensional approximating parameter spaces, which becomes richer and denser in the whole space with enlarging samples, allowing for consistent estimation for complex models (e.g., Grenander, 1981; Chen, 2007; Chen and Liao, 2015). DNN offer a “multi-layer” extension of the traditional sieve regression by modelling the connections among variables through data transformations from one layer to another (e.g., Fabozzi et al., 2019; Shen et al., 2023; Horel and Giesecke, 2020; Farrell et al., 2021). Such networks have a larger flexibility and potentially a faster convergence rate than the single-layer structures by increasing the sieve complexity to ensure consistent estimation while maintaining a relatively simple structure in each layer, and they are capable of improving on a larger scale with a growing sample size compared to traditional single-layer learning methods (e.g., Fabozzi et al., 2019; Shen et al., 2023; Husmeier, 2012; Farrell et al., 2021; Mathew et al., 2021). Husmeier (2012) demonstrated that a DNN architecture with multiple hidden layers and a larger scale of nodes contributes to the estimation of the conditional expectation function of the outcome. Such results were further studied by Chernozhukov et al. (2022) in bias estimation and bias reduction through de-noising the components of unobserved confounding factors from the outcome. Farrell et al. (2021) suggested to use DNNs to study the causal effect of a binary treatment without considering unobserved confounding in a semi-parametric model. However, in general, the theories and practices of incorporating DNNs in estimating omitted variable bias over unobserved confounding factors or examining how unobserved confounding factors alter the inference in a restricted model are still undeveloped.

In this chapter, we apply DNNs to derive unbiased estimators, for treatment effects subject to observed controls and unobserved confounders, in two scenarios. Scenario 1 is under a restricted condition, where we assume the unobserved part can be fully explained by a function of observed control variables; Scenario 2 is under a more general condition, where we assume the unobserved part can only be partially explained by observed controls and the remaining part is not correlated with the observed controls. We prove the consistency of the DNN-based estimators in both scenarios and derive the asymptotic normality for the estimator from Scenario 1, while the inference for Scenario 2 is performed by bootstrap. Essentially, the proposed method addresses the issue of omitted variable bias in observational studies by isolating the effects from the unobserved confounding factors using DNNs. To our best knowledge, such a method has not been developed in previous

studies.

The rest of this chapter is structured as follows. In section 2.2, we introduce our DNN-based estimators in both scenario 1 and scenario 2 and establish its large sample properties. In section 2.3, we illustrate the proposed method in a simulation analysis. Section 2.4 concludes.

2.2 Methodology and Asymptotics

In this section, we explore the estimation method and asymptotics in two scenarios. Scenario 1 assumes the unobserved part can be fully explained by a combination of linear and non-linear functions of the observed control variables. We first use an L_2 mean square error (MSE) function as the loss function to estimate the unobserved part using DNN and observed controls. Then, we estimate the treatment effect through a linear regression with a new outcome variable, created by removing the estimated unobserved part. We prove the consistency of the DNN-based estimators and derive their asymptotic normality.

In Scenario 2, we assume the unobserved part can only be partially explained by a function of observed control variables, and the remaining part is uncorrelated with observed controls. We apply a moment-condition-based loss function and estimate the conditional expectation outcome as a function of the treatment and observed controls, given the partial linear form of the conditional expectation function. We prove the consistency of the DNN-based estimators and the distribution of the estimator is provided by a bootstrap method.

Without loss of generality, we consider the situation of one endogenous treatment variable as an illustration throughout this section.

2.2.1 General Setup and Definitions

Consider a traditional omitted variable bias framework with a linear regression model,

$$Y = \beta X + \gamma^\top \omega + W + \epsilon, \tag{2.2.1}$$

where Y is a scalar outcome continuously distributed on $[-1, 1]$, X a scalar treatment continuously distributed on $[-1, 1]$, ω a d_ω -vector of observed controls continuously distributed on $[-1, 1]^{d_\omega}$, W an observable continuous scalar component, and ϵ the error term, such that

$$\mathbb{E}[\epsilon|X, \omega, W] = 0. \quad (2.2.2)$$

The treatment effect β cannot be estimated directly due to the unobservable part W , but it can be estimated through an alternative regression model. Specifically, consider

$$Y = \beta_s X + \gamma_s^\top \omega + \epsilon_s, \quad (2.2.3)$$

which is referred to as the “short” regression and (X, ω) the “short” list of regressors. Our goals are (1) to analyze the difference between the estimators in Equation 2.2.1 and Equation 2.2.3, and (2) to develop a consistent estimator for β from Equation 2.2.1 given only the “short” list of regressors (X, ω) .

In the following discussion, let $(Y_i, X_i, \omega_i, W_i)$ be an i.i.d. (independent and identically distributed) sample of the variables (Y, X, ω, W) for $i = 1, \dots, n$, where n is the sample size.

2.2.2 Scenario 1: Fully represented unobserved part

A fully representable unobserved part by the observable controls is assumed in this scenario.

Assumption 2.2.1. *The unobserved part W is a function of ω as such $W_i = f_0(\omega_i)$ for all $i = 1, \dots, n$, where $f_0 : [-1, 1]^{d_\omega} \rightarrow \mathbb{R}$ is a well-defined non-stochastic continuous function of an unknown form.*

Under assumption 2.2.1, the zero conditional mean assumption of the error term still holds as $\mathbb{E}[\epsilon_i|X_i, \omega_i, W_i] = \mathbb{E}[\epsilon_i|X_i, \omega_i, f_0(\omega_i)] = \mathbb{E}[\epsilon_i|X_i, \omega_i] = 0$ for all i . Assumption 2.2.1 could be approached when the observed controls are high-dimensional and have a strong explanatory power. A typical example is the study by (Card, 1999) on the causal effect of education on earnings, where ability is one of the unobserved variables. Assumption 2.2.1 in this empirical study indicates that if the number of observed variables is large enough,

immeasurable variable, such as ability, can be jointly described by a function of those observed controls, such as academic performance, social skills, creativity, stress tolerance, etc, and this function fully represents the relationship between the unobserved confounding and the observed controls.

Following the above setting, the regression function in Equation 2.2.1 can be rewritten as

$$Y_i = \beta X_i + \gamma^\top \omega_i + f_0(\omega_i) + \epsilon_i = g(A_i) + \epsilon_i, \quad (2.2.4)$$

subject to $\mathbb{E}[\epsilon_i|X_i, \omega_i] = \mathbb{E}[\epsilon_i|A_i] = 0$, where $g(A_i) = \beta X_i + f(\omega_i)$, with $f(\omega_i) := \gamma^\top \omega_i + f_0(\omega_i)$ and $A_i := (X_i, \omega_i)$, for all i . Note that β can be estimated by a basic least squares method if the function f were known – although f is not given, it can be obtained as $g(A)|_{X=0} = f(\omega)$ through the estimation of function g , where we incorporate DNNs.

Selecting a suitable DNN architecture can help enhance estimation, but the most appropriate DNN varies depending on the specific nature of each problem. To estimate the function g , we represent its network architecture with the Rectified Linear Unit (ReLU) activation function, as shown in Figure 2.1. The activation function ReLU is defined as

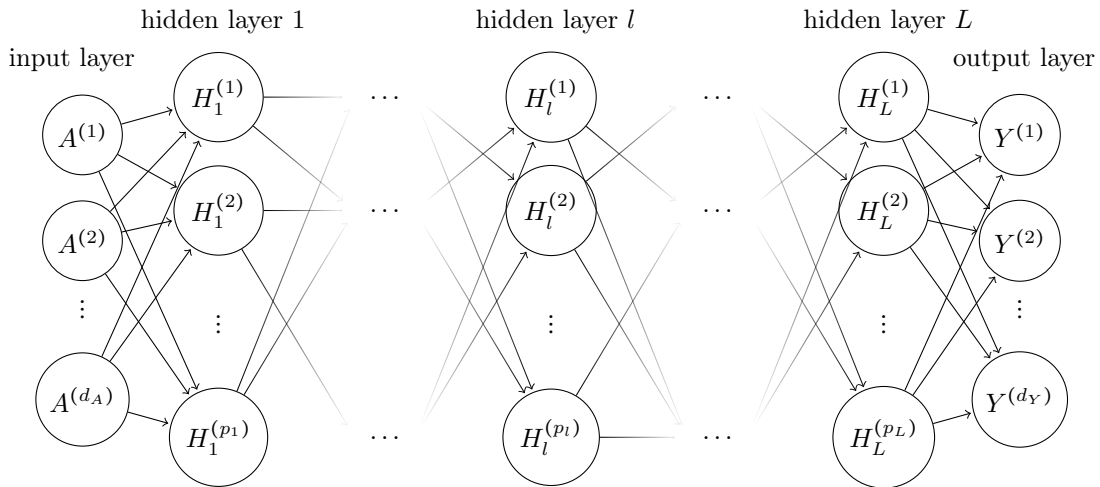


Figure 2.1: Deep Neural Network architecture with ReLU activation

$\text{ReLU}(a) = \max(a, 0)$, where $a \in \mathbb{R}$. A brief outline of DNN-ReLU construction was introduced by Anthony et al. (1999) and complete theory and applications are given by

Goodfellow et al. (2016). Essentially, ReLU is computationally efficient and usually has a fast convergence empirically.

As shown in Figure 2.1, the network can be represented by a function of the form

$$s(A) = Q_L \circ \text{ReLU} \circ \dots \circ Q_2 \circ \text{ReLU} \circ Q_1 \circ \text{ReLU} \circ Q_0 A. \quad (2.2.5)$$

The network has L hidden layers, and each layer has the corresponding width as noted in the set $\mathbf{p} := (p_0, p_1, \dots, p_L, p_{L+1})$, where p_l denotes the number of units in the l -th hidden layer for $l = 1, \dots, L$, and p_0 and p_{L+1} denote the numbers of the inputs and the outcomes, respectively, and thus, $p_0 = d_A$ and $p_{L+1} = 1$. We define the set of weight matrices $Q = (Q_0, \dots, Q_L)$, where Q_0 of dimension $p_1 \times p_0$ projects the input layer to the first hidden layer, Q_l of dimension $p_{l+1} \times p_l$ projects hidden layer l to $l + 1$ for $l = 1, \dots, L - 1$, and Q_L of dimension $p_{L+1} \times p_L$ projects the last hidden layer to the outcome layer. The space \mathcal{S} of all such networks is defined as

$$\mathcal{S}(L, \mathbf{p}) := \left\{ s \text{ takes the form of (2.2.5)} : \max_{l=0, \dots, L+1} \|Q_l\|_\infty \leq \infty \right\},$$

where $\|Q_l\|_\infty := \max_{1 \leq i \leq d_{Q,l+1}} \sum_{j=1}^{d_{Q,l}} |Q_{l,[i,j]}|$.

We denote a DNN representation for g with the network structure $s \in \mathcal{S}$ by $g_n^s(A; Q)$, which can be considered as a function of the set of variables A and the set of weight functions Q . Given the structure s , the estimators of the weights, denoted by $\hat{Q} := (\hat{Q}_0, \hat{Q}_1, \dots, \hat{Q}_L)$, can be obtained by minimizing the loss function

$$\text{Loss}_n^s(\tilde{Q}) := \frac{1}{n} \sum_{i=1}^n \{Y_i - g_n^s(A_i; \tilde{Q})\}^2, \quad (2.2.6)$$

which induces the estimated function $g_n^s(A; \hat{Q})$, and thus,

$$f_n^s(\omega; \hat{Q}) = g_n^s(A; \hat{Q})|_{X=0}. \quad (2.2.7)$$

Then, a DNN-based estimator $\hat{\beta}_{DNN}$ for β can be defined as such

$$\hat{\beta}_{DNN} := \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_i [Y_i - f_n^s(\omega_i; \hat{Q})] \right\}. \quad (2.2.8)$$

Note that, in practice, the optimal structure s can be selected by using data-driven meth-

ods, such as cross-validation.

Large sample properties: Consistency

Given that the function g can be approximated by a DNN, which can be estimated by $g_n^s(A_i; \hat{Q})$, the following assumptions are imposed to derive the asymptotic properties of the proposed DNN-based estimator.

Assumption 2.2.2. *With random samples $(Y_i, X_i, \omega_i)_i$, there exists an absolute constant $M > 0$ such that $\|\mathbb{E}[Y_i|X_i, \omega_i]\|_\infty \leq M$ for all i .*

As indicated by Assumption 2.2.2, the outcome is bounded and the covariates are continuously distributed, which is fairly common in non-parametric regression.

Assumption 2.2.3. *The function $g_\star(A) := \mathbb{E}[Y|X, \omega]$ lies in the Sobolev ball $W^{\beta_g, \infty}([-1, 1]^{d_c})$ of smoothness $\beta_g \in \mathbb{N}_+$ and $d_c := d_\omega + 1$, such that for all i , there is*

$$g_\star(A_i) \in W^{\beta_g, \infty}([-1, 1]^{d_c}) := \{h : \max_{\|\alpha\|_1 \leq \beta_g} \text{ess sup}_{A_i \in [-1, 1]^{d_c}} |D^\alpha h(A_i)| \leq C\}$$

for some constant $C > 0$, where $\alpha = (\alpha_1, \dots, \alpha_{d_c})$ is a vector of positive real numbers, $\|\alpha\|_1 = \alpha_1 + \dots + \alpha_{d_c}$, and $D^\alpha h$ is a weak derivative, as such $D^\alpha h(A) = \frac{\partial^{|\alpha|} h}{\partial A_1^{\alpha_1} \dots \partial A_{d_c}^{\alpha_{d_c}}}$.

This smoothness assumption is imposed to achieve the desired convergence speed in large sample properties.

Then with $\|\cdot\|_2$ denoting the l_2 -norm, we have Lemma 2.2.1 as follows.

Lemma 2.2.1. *Suppose Assumptions 2.2.2 and 2.2.3 hold. Let $\hat{g}_{DNN}(A_i) := g_n^s(A_i; \hat{Q})$ on $A_i \in [-1, 1]^{d_c}$ for all i be the deep ReLU network estimator defined by 2.2.5, for appropriate loss function in 2.2.6 with $H_n \asymp n^{\frac{d_c}{2(\beta_g + d_c)}} \log^2 n$ and $L_n \asymp \log n$, where H_n denotes the common order shared by all the width of all layers, and L_n is the number of hidden layers. Then with probability at least $1 - \exp(-n^{\frac{d_c}{\beta_g + d_c}} \log^8 n)$,*

$$\begin{aligned} \|\hat{g}_{DNN} - g\|_{L_2(A)}^2 &\leq C \cdot \left\{ n^{-\frac{\beta_g}{\beta_g + d_c}} \log^8 n + \frac{\log \log n}{n} \right\}, \\ \mathbb{E}[|\hat{g}_{DNN}(A) - g(A)|^2] &\leq C \cdot \left\{ n^{-\frac{\beta_g}{\beta_g + d_c}} \log^8 n + \frac{\log \log n}{n} \right\}, \end{aligned}$$

for a universal constant $C > 0$.

The proof of Lemma 2.2.1 is provided by Farrell et al. (2021). In the Lemma, the upper bound of the estimation error indicates that the optimal approximation power is reached only if the neural network architecture is optimal in terms of the depth L_n and the width H_n .

Lemma 2.2.2. *With Assumptions 2.2.1, 2.2.2 and 2.2.3, and Lemma 2.2.1, there exists some network s such that $\mathbb{E}[\|\hat{g}_{DNN} - g\|_{L_2(A)}^2] \rightarrow 0$ as $n \rightarrow \infty$.*

The upper bound of the estimation error of \hat{g}_{DNN} is provide in Lemma 2.2.1. With an appropriate neural network architecture, this upper bound converges to 0 as $n \rightarrow \infty$, which implies that the estimation error of \hat{g}_{DNN} will also converge to 0 as $n \rightarrow \infty$, as shown in Lemma 2.2.2.

Lemma 2.2.3. *Let $\hat{f}_{DNN}(\omega_i) := f_n^s(\omega_i; \hat{Q})$ on $\omega_i \in [-1, 1]^{d_\omega}$ for all i . With Equation 2.2.1 and Lemma 2.2.2, there exists some network s such that $\mathbb{E}[\|\hat{f}_{DNN} - f\|_{L_2(\omega)}^2] \rightarrow 0$ as $n \rightarrow \infty$.*

Given $f(\omega_i)$ as the non-parametric component of $g(A_i)$, the convergence of the estimation error of \hat{f}_{DNN} also holds as $n \rightarrow \infty$, which further implies the consistency of $\hat{\beta}_{DNN}$.

Assumption 2.2.4. *For all $i = 1, \dots, n$*

(a). *there exists a strictly positive-definite real-valued matrix Σ_ϵ , such that*

$$\Sigma_\epsilon := \text{Var}(\epsilon_i | X_i, \omega_i);$$

(b). *there exists a real value $\Sigma_{XX} := \mathbb{E}[X_i^2]$ for all i .*

Theorem 2.2.1. *Suppose Assumptions 2.2.1 to 2.2.4 hold. With $f_n^s(A_i; \hat{Q})$ satisfying Lemma 2.2.3, and $\hat{\beta}_{DNN}$ as in (2.2.8), there is $\|\hat{\beta}_{DNN} - \beta\|_2^2 \xrightarrow{p} 0$.*

The consistency of $\hat{\beta}_{DNN}$ holds if $n^{-1} \sum_{i=1}^n X_i \epsilon_i \xrightarrow{p} 0$ and $n^{-1} \sum_{i=1}^n X_i X_i^\top$ is well-defined and invertible as $n \rightarrow \infty$, which are justified by Equation 2.2.2 and Assumption 2.2.4(b) respectively, under some \hat{f}_{DNN} that satisfies Lemma 2.2.3, among other suitable conditions.

Large sample properties: Asymptotic normality

Under proper conditions, the asymptotic normality can be derived under the high-dimensional central limit theorem (CLT) by Chernozhukov et al. (2017), with a desirable convergence rate of the DNN estimated function \hat{f}_{DNN} .

Assumption 2.2.5. For all $i = 1, \dots, n$

(a). there exists some $\delta > 0$, such that $\mathbb{E}[\epsilon_i^{2+\delta}] < \infty$, $\mathbb{E}[\|X_i \epsilon_i\|^{2+\delta}] < \infty$;

(b). the real value $\Omega = \mathbb{E}[X_i^2 \epsilon_i^2]$ exists;

(c). there exists some $M > 0$, such that $\mathbb{E}[\epsilon_i^2 | X_i, \omega_i] < M < \infty$.

The additional Assumption 2.2.5 are required for the central limit theorem (CLT). Then we have the following theorem for the asymptotic normality.

Theorem 2.2.2. Suppose Assumptions 2.2.1 to 2.2.5 hold. With \hat{f}_{DNN} satisfying Lemma 2.2.3, and $\hat{\beta}_{DNN}$ as in (2.2.8),

$$\sqrt{n}(\hat{\beta}_{DNN} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_{XX}^{-2} \Omega \Sigma_{XX}^{-1}).$$

2.2.3 Scenario 2: Partially represented unobserved part

For this scenario, we develop a consistent estimator of β under a more general condition, assuming that the unobserved part can only be partially explained by the observed control variables and the remaining part is uncorrelated with observed controls.

Assumption 2.2.6. The unobserved part W can be written as $W_i = f_0(\omega_i) + M_i$ for all i . $f_0(\omega_i)$ denotes the part that can be represented by the observed controls, where $f_0 : [-1, 1]^{d_\omega} \rightarrow \mathbb{R}$ is a well-defined non-stochastic continuous function of an unknown form. The remaining part M_i is such that $\mathbb{E}[M_i | \omega_i] = 0$, $\text{corr}(M_i, X_i) \neq 0$, and $\mathbb{E}[\epsilon_i | M_i] = 0$ for all i .

Under assumption 2.2.6, the zero conditional mean assumption of the error term still holds as $\mathbb{E}[\epsilon_i|X_i, \omega_i, W_i] = \mathbb{E}[\epsilon_i|X_i, \omega_i, f_0(\omega_i), M_i] = \mathbb{E}[\epsilon_i|X_i, \omega_i, M_i] = 0$. Assumption 2.2.6 could be met in practice when there exists a decomposition of the unobserved part W , partitioning it into two components: one that can be fully represented by the observed controls (i.e., f_0), and another that cannot be represented but is orthogonal to all functions of the observed controls (i.e., M). The assumption on M is reasonable, because if it were not orthogonal to some functions of the observed controls, that relationship would have been picked as a part of f_0 . As M is correlated with the treatment, the existence of M still alter the inference in the restricted model.

We rewrite the regression in 2.2.1 as

$$Y_i = \beta X_i + \gamma^\top \omega_i + f_0(\omega_i) + M_i + \epsilon_i = g[(X_i, \omega_i)] + M_i + \epsilon_i, \quad (2.2.9)$$

where $g[(X_i, \omega_i)] = \beta X_i + f(\omega_i)$, and we re-define $f(\omega_i) := \gamma^\top \omega_i + f_0(\omega_i)$. Then, using a similar system of notations as in Scenario 1, a semi-parametric network is constructed as follow:

$$s(X_i, \omega_i) = \beta X_i + Q_L \circ \text{ReLU} \circ \dots \circ Q_2 \circ \text{ReLU} \circ Q_1 \circ \text{ReLU} \circ Q_0 \omega_i. \quad (2.2.10)$$

The network has L hidden layers, and each layer has the corresponding width as noted in the set $\mathbf{p} := (p_0, p_1, \dots, p_L, p_{L+1})$, whereas in this case, $p_0 = d_\omega$, and the space \mathcal{S} of all such networks is defined as

$$\mathcal{S}(L, \mathbf{p}) := \left\{ s \text{ takes the form of (2.2.10)} : \max_{l=0, \dots, L+1} \|Q_l\|_\infty \leq \infty, |\beta| \leq \infty \right\},$$

where $\|Q_l\|_\infty := \max_{1 \leq i \leq d_{Q_l, l+1}} \sum_{j=1}^{d_{Q_l, l}} |Q_{l, [i, j]}|$.

We denote a DNN representation for g by $g_n^s[(X_i, \omega_i); \beta, Q] := \beta X + f_n^s(\omega_i; Q)$, with the sample size n , the network structure $s \in \mathcal{S}$, and weights Q . Then the estimators of the parameters $(\hat{Q}, \hat{\beta}) := (\hat{Q}_0, \hat{Q}_1, \dots, \hat{Q}_L, \hat{\beta})$ can be obtained by minimizing the loss function

$$\text{Loss}_n^s(\tilde{Q}, \tilde{\beta}) := \left\{ \frac{1}{n} \sum_{i=1}^n [(Y_i - \tilde{\beta} X_i - f_n^s(\omega_i; \tilde{Q})) \times k(\omega_i)] \right\}^2, \quad (2.2.11)$$

where $(\tilde{Q}, \tilde{\beta})$ denotes any estimator of the corresponding parameter, and $k(\omega_i)$ is any well-defined non-stochastic function as such $k : [-1, 1]^{d_\omega} \rightarrow \mathbb{R}$. It is worth noting that such a loss function is a moment-condition-based loss function. As mentioned by Bennett et al. (2019), under certain conditions, the loss function based on the true moment conditions is also an appropriate loss function in DNN architecture to obtain an estimated function, which converges to the underlying function as the sample size grows towards infinity.

With the estimated parameters $(\hat{Q}, \hat{\beta})$ and any non-zero real-valued number r , we can define a DNN-based estimator $\hat{\beta}_{DNN}$ for β , such that

$$\hat{\beta}_{DNN} = \frac{1}{nr} \sum_{i=1}^n \{g_n^s[(X_i + r, \omega_i); \hat{Q}, \hat{\beta}] - g_n^s[(X_i, \omega_i); \hat{Q}, \hat{\beta}]\}. \quad (2.2.12)$$

Large sample properties: Consistency

We state the following assumption on the set \mathcal{K} of $k(\omega_i)$ used in the loss function 2.2.11.

Assumption 2.2.7. \mathcal{K} is a set of function, with a convex closure $\bar{\mathcal{K}}$, such that for all $k \in \bar{\mathcal{K}}$, there exists some $\delta > 0$ that satisfies $\|\mathbb{E}[(Y_i - \tilde{\beta}X_i - f(\omega_i))k(\omega_i)]\| \leq \delta \implies \|\mathbb{E}[Y_i - \tilde{\beta}X_i - f(\omega_i)|\omega_i]\| \leq C\delta$ for all i and some constant $C > 0$.

Assumption 2.2.7 provides some restrictions on function k to ensure that if the expectation $\mathbb{E}[(Y_i - \tilde{\beta}X_i - f(\omega_i))k(\omega_i)]$ is small enough or equal to zero for function k , then the conditional expectation $\mathbb{E}[Y_i - \tilde{\beta}X_i - f(\omega_i)|\omega_i]$ will also be small enough or equal to zero. Then we have the following Lemma for the estimator.

Lemma 2.2.4. Suppose Assumptions 2.2.2, 2.2.3 and 2.2.7 hold. Let $\hat{g}_{DNN}[(X_i, \omega_i)] := g_n^s[(X_i, \omega_i); \hat{Q}, \hat{\beta}]$ be a deep ReLU network estimator as in 2.2.10 derived from the loss function in 2.2.11, with $H_n \asymp n^{\frac{d_c}{2(\beta_g + d_c)}} \log^2 n$ and $L_n \asymp \log n$. Then with probability at least $1 - \exp(-n^{\frac{d_c}{\beta_g + d_c}} \log^8 n)$,

$$\|\hat{g}_{DNN} - g\|_{L_2(X, \omega)}^2 \leq C \cdot \left\{ n^{-\frac{\beta_g}{\beta_g + d_c}} \log^8 n + \frac{\log \log n}{n} \right\},$$

for a universal constant $C > 0$.

Supported by Farrell et al. (2021) and Lewis and Syrgkanis (2018), with an optimal neural network architectures of H_n and L_n and some k satisfying Assumption 2.2.7 in the loss function, the upper bound between estimated function and underlying function with the optimal approximation power is provided in Lemma 2.2.4.

Lemma 2.2.5. *With Assumptions 2.2.2, 2.2.3, 2.2.6 and 2.2.7, and Lemma 2.2.4, there exists some network s such that $\mathbb{E}[\|\hat{g}_{DNN} - g\|_{L_2(X,\omega)}^2] \rightarrow 0$ as $n \rightarrow \infty$.*

Then we have the following theorem on consistency.

Theorem 2.2.3. *Suppose Assumptions 2.2.2 to 2.2.7 hold. With $g_n^s[(X_i, \omega_i); \hat{Q}, \hat{\beta}]$ satisfying Lemma 2.2.5, and $\hat{\beta}_{DNN}$ as in (2.2.12), there is $\|\hat{\beta}_{DNN} - \beta\|_2^2 \xrightarrow{P} 0$.*

Large sample properties: Inference

Given the consistency of the estimator $\hat{\beta}_{DNN}$, the asymptotic distribution could be generated by a bootstrap approach. We present numerical evidence for this large sample property in the simulation study in Section 2.3. We simulate data with the relationship in Scenario 2 with different data generating processes. The distributions of estimator $\hat{\beta}_{DNN}$ appear normal and a Kolmogorov–Smirnov (KS) test does not reject normality. In this study, we did not derive the analytical asymptotic distribution of the estimator $\hat{\beta}_{DNN}$ for scenario 2, due to that the computation has some extra steps that may cause interactions among error components from the semi-parametric DNN training, and further, lead to some ambiguity in the leading error. However, the numerical results provide some hint for further exploring the analytical asymptotic distribution of this estimator in further studies. Nevertheless, in practice, a bootstrap approach could be an acceptable way to calculate standard errors and generate asymptotic distribution of $\hat{\beta}_{DNN}$ for inference.

2.3 Simulation Analysis

We now use a simulation study to illustrate the proposed DNN-based estimation and its asymptotics in both two scenarios.

2.3.1 Simulation Study for Scenario 1

Consider the regression model as explained above in Equation 2.2.4,

$$Y_i = \beta X_i + \gamma^\top \omega_i + W_i + \epsilon_i = \beta X_i + \gamma^\top \omega_i + f_0(\omega_i) + \epsilon_i.$$

The variables and the parameters are generated as follows, such that for $i = 1, \dots, n$,

$$\omega_i \sim \mathcal{N}(0, \mathbf{I}_5), X_i, \epsilon_i \sim \mathcal{N}(0, 1), \text{ and } \text{corr}(X_i, \omega_i^{(j)}) = 0.05 \text{ for } j = 1, 2, \dots, 5.$$

and

$$\beta = 1, \gamma = (1, 1, 1, 1, 1)^\top, \text{ and } f_0(\omega) = 2 \sum_{j=1}^5 [\omega^{(j)}]^\frac{1}{3}.$$

Sample of sizes $n = 1000, 2000$ are generated repeatedly for $S = 1000$ times based on the above data generating process (DGP). With $W_i := f_0(\omega_i)$ being unobserved, we consider a restricted regression model explained in 2.2.3 and obtain the OLS estimator $\hat{\beta}_{OLS}$. The estimator $\hat{\beta}_{OLS}$ is biased as $\text{cov}(X_i, W_i) = \text{cov}(X_i, f_0(\omega_i)) \neq 0$. In this design, we compare the performance of the proposed DNN-based estimator with the restricted OLS estimator to demonstrate that the proposed estimator can help correct this bias by capturing the potential non-linear relationships between observed and unobserved variables. To calculate the function f_0 , we use the real value of the power $\frac{1}{3}$.

In this simulation analysis, we use a ReLU activation function. As for the DNN structure, in this simulation study, we employ a network with two hidden layers, where the first hidden layer includes 200 units and the second includes 100. To obtain a more thorough search for the optimal DNN structure in practice, one can make it data-driven from a larger pool of candidates as needed by the complexity of the problems. To train the DNN under any given structure, we split the total sample of n into 80% of training sample and 20% of validation sample, where the training sample is to obtain a $g_n^s(A_i; \hat{Q})$ by minimizing the loss function, and the validation sample is to monitor the out-of-sample performance for possible over-fitting. After a sufficiently large number of iterations, the trained model from the iteration with the smallest validation loss will be selected for the estimated function $g_n^s(A_i; \hat{Q})$. Then we can move on to obtain the estimated coefficient $\hat{\beta}_{DNN}$ by using (2.2.7).

The 1000 rounds of simulation yield 1000 values for the restricted OLS estimator and

1000 of the proposed DNN-based estimator. We compute the empirical mean, bias, standard error (SE), and mean squared error (MSE) for both estimators, and the relative bias of the DNN-based estimator to the OLS estimator. We repeat the simulation at different sample sizes $n = 1000, 2000$. The results are displayed in Table 2.1.

Table 2.1: Scenario 1: Estimation Statistics

| n=1000 | Mean | Bias | SE | MSE | Relative bias to OLS |
|----------------|--------|--------|--------|--------|----------------------|
| OLS-Restricted | 1.4585 | 0.4585 | 0.0493 | 0.2127 | 100% |
| DNN | 1.0147 | 0.0147 | 0.0285 | 0.0010 | 3.21% |
| n=2000 | Mean | Bias | SE | MSE | Relative bias to OLS |
| OLS-Restricted | 1.4495 | 0.4495 | 0.0358 | 0.2033 | 100% |
| DNN | 1.0133 | 0.0133 | 0.0266 | 0.0008 | 2.96% |

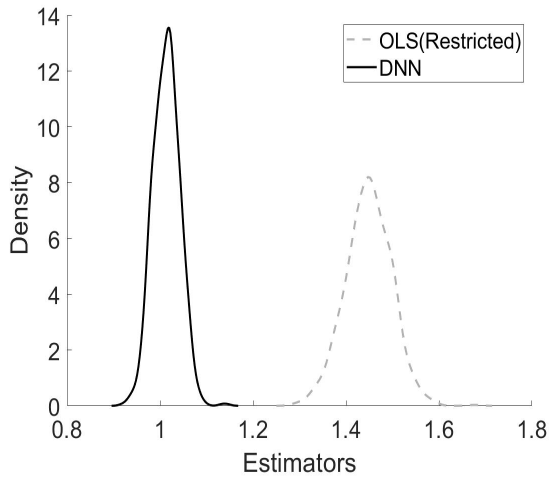
Under the sample size $n = 1000$, as shown in Table 2.1, the relative bias of the DNN-based estimator is around 3.21% and the MSE of the DNN-based estimator are 0.0010, which are both much smaller than the restricted OLS estimator. A comparison among their distributions is presented in Figure 2.2(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 2.2(b). The p value for the KS test is 0.3481, which means we can not reject that the distribution of the DNN-based estimators is normal.

Under the sample size $n = 2000$, as shown in Table 2.1, the relative bias of the DNN-based estimator is around 2.96% and the MSE of the DNN-based estimator are 0.0008, which are both much smaller than the restricted OLS estimator. A comparison among their distributions is presented in Figure 2.3(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 2.3(b). The p value for the KS test is 0.4562, which means we can not reject that the distribution of the DNN-based estimators is normal.

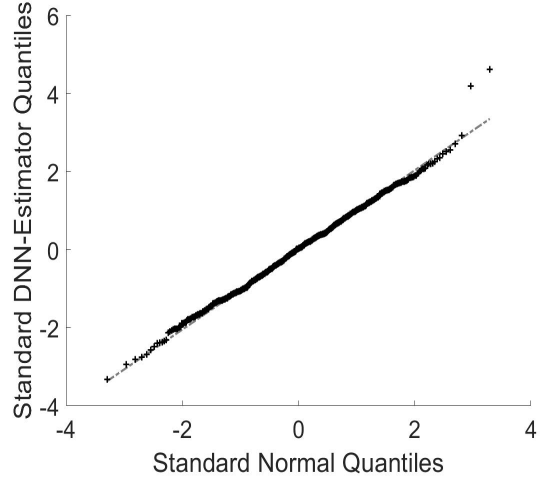
2.3.2 Simulation Study for Scenario 2

Consider the regression model as explained above in 2.2.9 for Scenario 2,

$$Y_i = \beta X_i + \gamma^\top \omega_i + f_0(\omega_i) + M_i + \epsilon_i = g[(X_i, \omega_i)] + M_i + \epsilon_i,$$

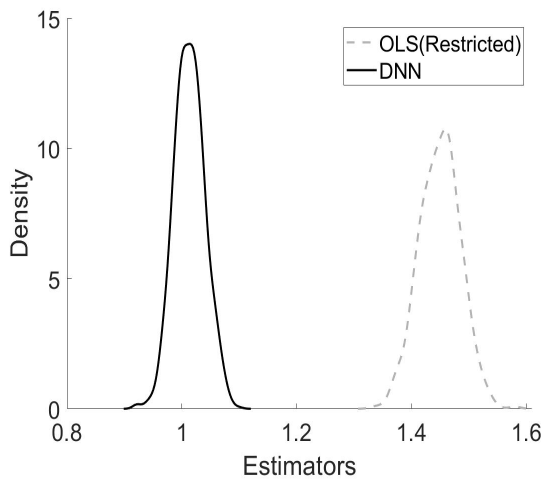


(a) A comparison among estimators

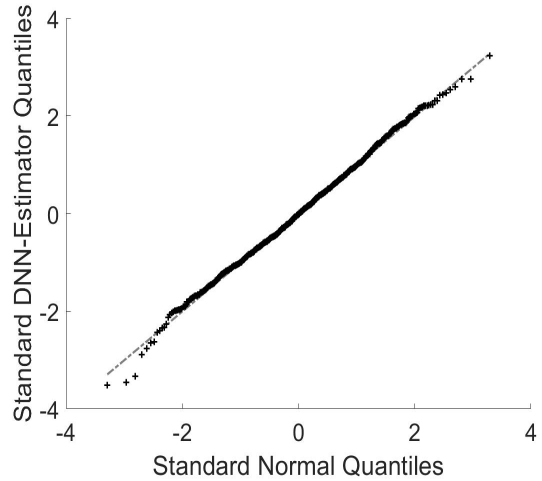


(b) $\hat{\beta}_{\text{DNN}}$ -Normal Quantile-Quantile plot

Figure 2.2: Distributions of the OLS and DNN Estimator in Scenario 1, $n = 1000$



(a) A comparison among estimators



(b) $\hat{\beta}_{\text{DNN}}$ -Normal Quantile-Quantile plot

Figure 2.3: Distributions of the OLS and DNN Estimator in Scenario 1, $n = 2000$

The variables and the parameters are generated as follows, such that for $i = 1, \dots, n$,

$$\omega_i \sim \mathcal{N}(0, \mathbf{I}_5), X_i, \epsilon_i \sim \mathcal{N}(0, 1), \text{ and } M_i \sim \mathcal{N}(0, 4), \text{ with}$$

$$\text{corr}(X_i, \omega_i^{(j)}) = 0.05 \text{ for } j = 1, \dots, 5, \text{ and } \text{corr}(X_i, M_i) = 0.2,$$

and

$$\beta = 1, \gamma = (1, 1, 1, 1, 1)^\top, \text{ and } f_0(\omega) = 2 \sum_{j=1}^5 [\omega^{(j)}]^\frac{1}{3}.$$

Similar to Scenario 1, samples of sizes $n = 1000, 2000$ are generated repeatedly for $S = 1000$ times based on the above DGP. We still compare the performance of the proposed DNN-based estimator with the restricted OLS estimator to demonstrate that the proposed estimator can help correct this bias even though the unobserved part can only be partially captured by the observed controls, where the exogeneity of the remaining part M towards the observed controls plays an essential role. To calculate the function f_0 , we use the real value of the power $\frac{1}{3}$.

In this simulation analysis, we use a two-hidden-layer DNN with the ReLU activation function, where the first hidden layer includes 250 units and the second includes 125. To train for the estimated function $g_n^s[(X_i, \omega_i); \hat{Q}, \hat{\beta}]$ under any given structure, we split the total sample of n into 75% of training sample and 25% of validation sample. After a sufficiently large number of iterations, the trained model from the iteration with the smallest validation loss will be selected for the estimated function $g_n^s[(X_i, \omega_i); \hat{Q}, \hat{\beta}]$. Then we can move on to obtain the estimated coefficient $\hat{\beta}_{\text{DNN}}$ by using (2.2.12). The choice of r is based on the empirical observations of X_i , so that the value of $X_i + r$ remains in the domain of the function g for all $i = 1, \dots, n$. Practically, there can be various ways determine such a value for r . The corresponding results, similar to those presented in Table 2.1, are shown in Table 2.2

Table 2.2: Scenario 2: Estimation Statistics

| n=1000 | Mean | Bias | SE | MSE | Relative bias to OLS |
|----------------|--------|--------|--------|--------|----------------------|
| OLS-Restricted | 1.6730 | 0.6730 | 0.0600 | 0.4566 | 100% |
| DNN | 1.0295 | 0.0295 | 0.0531 | 0.0037 | 4.38% |
| n=2000 | Mean | Bias | SE | MSE | Relative bias to OLS |
| OLS-Restricted | 1.6783 | 0.6783 | 0.0410 | 0.4618 | 100% |
| DNN | 1.0299 | 0.0299 | 0.0518 | 0.0036 | 4.41% |

Under the sample size $n = 1000$, as shown in Table 2.2, the relative bias of the DNN-based estimator is around 4.38% and the MSE of the DNN-based estimator are 0.0037,

which are both much smaller than the restricted OLS estimator. A comparison among their distributions is presented in Figure 2.4(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 2.4(b). The p value for the KS test is 0.2789, which means we can not reject that the distribution of the DNN-based estimators is normal.

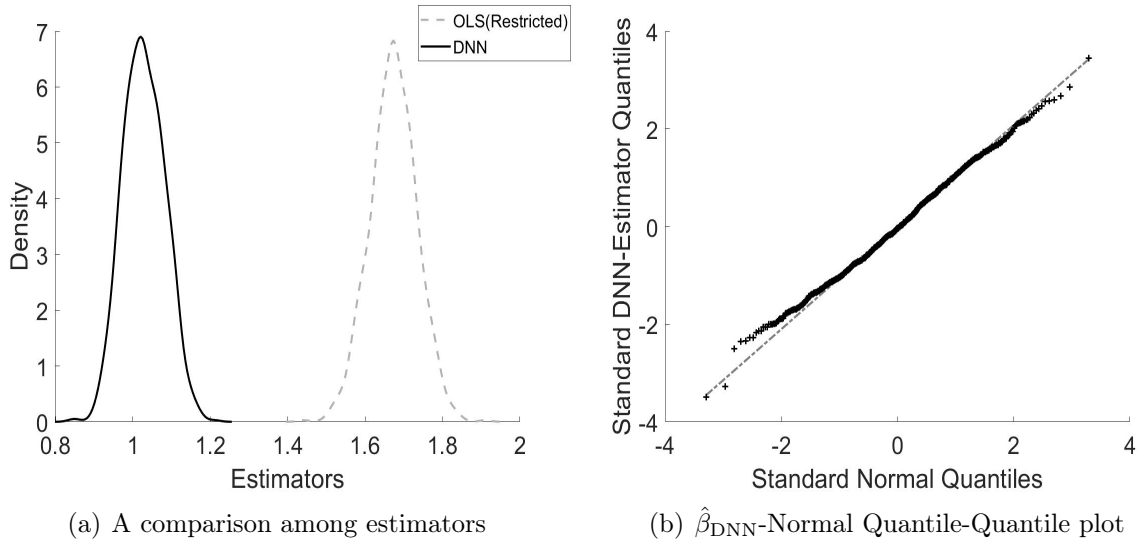


Figure 2.4: Distributions of the OLS and DNN Estimator in Scenario 2, $n = 1000$

For this scenario, we do not see a notable improvement as the sample size grows – although the MSE decreases from $n = 1000$ to $n = 2000$, the relative bias of the DNN-based estimator under $n = 2000$ is slightly larger than that under $n = 1000$. However, the asymptotic normality appears to be robust. A comparison among their distributions is presented in Figure 2.5(a), and a Q-Q plot of the standard normal against the DNN-based estimator standardized by the empirical mean and variance across simulations is shown in Figure 2.5(b). The p value for the KS test is 0.2408, which means we can not reject that the distribution of the DNN-based estimators is normal.

The empirical sampling distribution of $\hat{\beta}_{DNN}$ in this section appears normal according to the QQ-plot and fails to reject the KS normality test. This result provides some hint for further exploring the analytical asymptotic distribution of this estimator in further studies. Nevertheless, in practice, a bootstrap approach could be an acceptable way to calculate

standard errors and generate asymptotic distribution of $\hat{\beta}_{\text{DNN}}$ for inference.

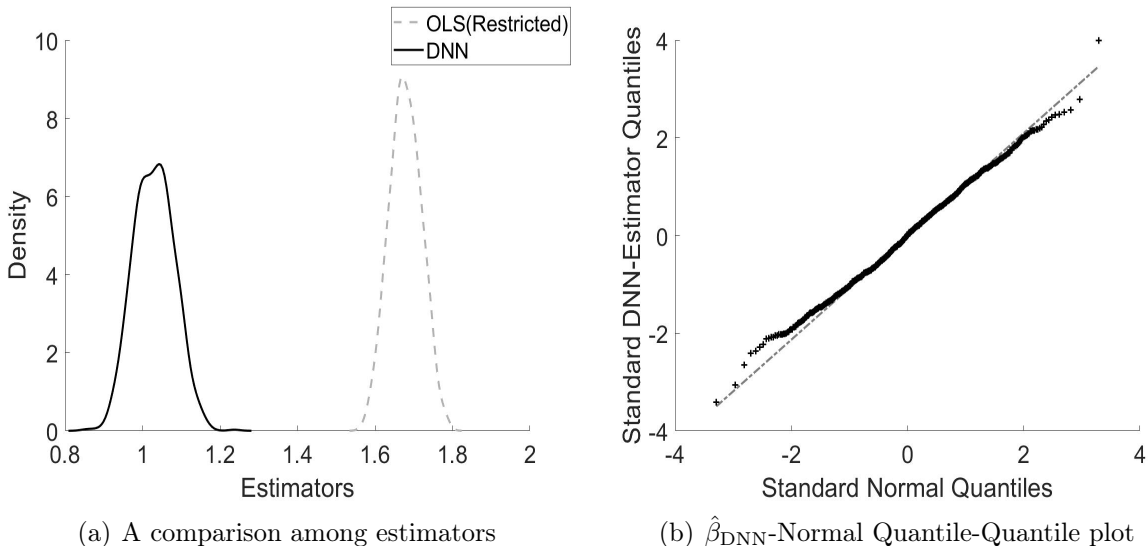


Figure 2.5: Distributions of the OLS and DNN Estimator in Scenario 2, $n = 2000$

2.4 Conclusion

Omitted variable bias is a significant concern in causal inference. In this chapter, we tackle a linear causal effect of a continuous treatment variable in a regression model subject to unobservable confounders.

To achieve reliable bias-adjusted inference, we incorporate DNNs to develop consistent estimators to the treatment effect in two scenarios. Scenario 1 is under restricted conditions, where we assume the unobserved part can be fully explained by a function of observed control variables. In this scenario, all the significant controls are considered to be captured by the observables in some formats, and thus, no unobserved confounder is left behind uncontrolled. Scenario 2, however, is under more general conditions, where we assume the unobserved part can be partially explained by control variables and the remaining part is uncorrelated with observed controls. In this scenario, a part of the unobserved confounder is left uncontrolled, and some constraints will be necessary to regulate this confounder,

whence we assume that it is orthogonal to all functions of the observable controls and is exogenous towards the error term. Such constraints induce moment conditions, based on which the parameters can be estimated. Essentially, the proposed method addresses omitted variable bias by isolating the effects of unobserved confounders using DNNs. We derive the asymptotic properties of the DNN-based estimator and demonstrate its performance in a simulation study.

It is important to note that the proposed methods focus exclusively on linear treatment effects and cases where treatments are additively separable from the controls and confounders. If there are more complex interactions between the treatments and the controls or the confounders, the method will require further generalization or modification.

Chapter 3

Test for High Dimensionality of Random Vectors¹

3.1 Introduction

Prior to 1950, most of practical problems consisted of a relatively large number of experimental units with a relatively small number of features which were measured (Rowell and Walters, 1976). Therefore, traditional theories and practice were limited to the “small dimension of variables and large sample size” scenario. This scenario naturally reflected the contemporary limitations of computers and graphical display. Over the last 25 years, however, Lindsay et al. (2004) pointed out that the environment for practical problems has changed dramatically, with the huge evolution of data acquisition technologies and computing facilities. The main scenarios to be investigated steadily evolve into the “large dimension and small sample size”, or in some cases “large dimension and large sample size”. With the latest development of computing techniques, such as the deep neural networks, this allows for data with much larger dimensions to be dealt with. Most of the latest large language models have contained more than 100 billions trainable parameters in Zhao et al. (2023). Not only do the techniques develop rapidly, but the high-dimensional theories also advance dramatically. Theoretical studies have focused on two aspects: investigate new

¹This chapter is co-authored with Tao Chen.

theories under high dimension scenarios and modify classical theories.

An increasing number of novel and useful properties are found under the high-dimensional scenario. Specifically, according to probability theory, it is a fundamental regularity that some representative laws exist in averaging many “individual” dimensions (Donoho et al., 2000). For instance, the concentration of measure phenomenon appears, first introduced by Gromov and Milman (1983), which means that random fluctuations from a Lipschitz function on the high dimensional sphere are bounded, and the tails behave no worse than a standard normal distribution in the tails. It is extend to the theories of extreme-value distributions of high-dimensional random vectors, where, for example, the maximum or the extreme values of Euclidean norm of D -dimensional Gaussian random variables have a limiting distribution when D approaches infinity (e.g., Fisher and Tippett, 1928; Donoho et al., 2000). Additionally, Vershynin (2020) provided many practical theories for random vectors and matrices, such as estimating concentration of the norm, approximating isometries, etc. These theories only work when the number of coordinates of random vectors and the entries of random matrices are sufficiently large. Specifically, when the dimension of random vectors or matrices grows increasingly, some good properties start to appear.

On the other hand, modification of the classical theories started, since many classical methods will fail if the dimension is sufficiently large compared to the sample size. This type of failure was firstly noticed by Dempster (1958), who showed that the classical Hotelling’s T-squared test became undefined as the number of variables became close to, or even exceeded the number of degrees of freedom within sample for estimation of the variance and co-variance matrix. A related simulation, which showed the failure of the Hotelling’s T-squared test, was also proposed by Bai and Saranadasa (1996). On the other hand, some classical tests for random vectors need to be re-built as their properties have changed under the high-dimensional scenario (e.g., Huber, 1973; Portnoy, 1984; Portnoy, Stephen, 1985; Portnoy, Stephen, 1988; He and Shao, 2000). For instance, the classical F-test and likelihood-ratio test, which will fail if the dimension is large compared to the sample size, which are corrected by Calhoun (2011) and Sur et al. (2019), respectively. Although the “curses of dimensionality” was pointed out by Bellman and Kalaba (1957) and the “blessings of dimensionality” are less widely noted, the “high dimensions” seems no longer just a “curse”, but rather a “gift” that can be utilized in modern theories.

The high dimensionality of random vectors, without any clear definition in the liter-

ature, is defined, in this chapter, as the scenario in which the high-dimensional theories hold or the high-dimensional properties appear. The idea behind the high dimensionality is comparable to what behind the central limit theorem (CLT), in which the normal distribution appears in the case of a sufficiently large sample size. In other words, under certain assumptions, statistical and probabilistic methodology that works for normal distributions can also be applicable to problems involving other types of distributions as the sample size approaches infinity. Such applicability sometimes collapses, given a finite sample in practice. In general, sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold. However, Lehmann (1999) showed the distribution of the arithmetic mean of independent random variables from a binomial distribution $B(N, p)$ is still not a satisfactory approximation of a normal distribution even when N is larger than 90, where the parameter p is equal to 0.05 and N is the sample size. It is also stated that the speed of convergence is dependent on the underlying distribution of the sample. Therefore, the general judgment criteria for an adequate sample size is not always reliable, which constitutes a major concern for the failure of CLT. Determining how large the sample size N is adequate to hold the CLT is essential in the finite sample. Similarly, one might ask how large the dimension D of random vectors need to be for these high-dimensional theories to hold. This is what we aims to answer throughout our proposed test for high dimensionality.

Compared to the finite or low-dimensional scenario, the high-dimensional scenario becomes more common, which we are facing in reality but we usually do not realize it. Although many methods have been proposed for high-dimensional scenario; however, a more basic question about whether random vectors in a finite data sample is in high dimensionality or not, has not been considered in the literature. In other words, there is no method which determines whether the high-dimensional theories can be applied for random vectors in a finite data sample. Discussing a global threshold between the high and non-high dimensional scenarios for random vector is also essential and need to be explored. This chapter proposes a general testing method to distinguish high dimensionality of random vectors from non-high. The null hypothesis in this chapter is defined as

H_0 : The dimension of the random vector is high.

Failure to reject the null hypothesis indicates that the random vectors are in high dimensions; and the high-dimensional theories can be applied. If the null hypothesis was rejected,

the random vectors fall into the scenario of undefined situation in which the availability of both classical and high-dimensional theories is unknown. Additionally, we provide guidance to determine the thresholds in the test for high dimensionality of random vectors based on a Monte Carlo study.

In section 3.2, the test for high dimensionality of random vectors is presented. Section 3.3 shows the implementation of the proposed tests for random vectors. A Monte Carlo study is presented in section 3.4.

Notation. N and D represent the sample size and the dimension of random vectors. D is allowed to approach N proportionally, but is less than N . We work on the random vectors $\{\mathbf{V}_n\}_{n=1}^N$ whose realization is $\{\mathbf{v}_n\}_{n=1}^N$. Besides, the ℓ_2 -norm is denoted by $\|\cdot\|$.

3.2 Tests of Random Vectors

In this section, we first present the test statistic for multivariate normal random vectors and its asymptotic property, and then extend the results to general random vectors.

3.2.1 Multivariate Normal Vectors

Identity Covariance Matrix

We first focus on the i.i.d. (independent and identically distributed) multivariate normal random vectors. Suppose there is a sequence of i.i.d. D -dimensional random vectors $\{\mathbf{V}_n\}_{n=1}^N$ such that $\mathbf{V}_n \sim \mathcal{N}(0, I_{D \times D})$ for each n . Our test statistic is defined as the following:

$$T_1 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \mathbf{V}_n \right) \right\|^2 - 1 \right). \quad (3.2.1)$$

Theorem 3.2.1. $T_1 \xrightarrow{d} \mathcal{N}(0, 1)$ as $N, D \rightarrow \infty$.

Proof: It is easy to show that $T_1 \xrightarrow{d} \mathcal{N}(0, 1)$ as N and $D \rightarrow \infty$. Rewrite $\left\| \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \mathbf{V}_n \right) \right\|^2$ as $\sum_{d=1}^D \left(\frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{V}_{n,d} \right)^2$ where $\mathbf{V}_{n,d}$ is the d th element of the vector \mathbf{V}_n . Since we know

that $\frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{V}_{n,d} \sim \mathcal{N}(0, 1)$ is independent of N , and $\frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{V}_{n,d}$ is independent w.r.t. index d , with $\mathbb{E} \left[\left(\frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{V}_{n,d} \right)^2 \right] = 1$ and $\text{var} \left[\left(\frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{V}_{n,d} \right)^2 \right] = 2$, it implies that the central limit theorem gives us

$$T_1 \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } N, D \rightarrow \infty.$$

□

Non-Identity Matrix

Now, we consider that the covariance matrix of the random vector is no longer an identity covariance matrix and release the assumption of i.i.d.. Suppose the covariance matrix of \mathbf{V}_n is Ω_D , i.e., $\mathbf{V}_n \sim \mathcal{N}(0, \Omega_D)$ for each n . Standardize \mathbf{V}_n by

$$\tilde{\mathbf{V}}_n := \Omega_D^{-\frac{1}{2}} \mathbf{V}_n \sim \mathcal{N}(0, I_{D \times D}).$$

Our test statistic is defined as the following:

$$T_2 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \Omega_D^{-\frac{1}{2}} \mathbf{V}_n \right) \right\|^2 - 1 \right). \quad (3.2.2)$$

Theorem 3.2.2. $T_2 \xrightarrow{d} \mathcal{N}(0, 1)$ as $N, D \rightarrow \infty$.

This is identical to the case of multivariate normal vectors with identity covariance matrix in Theorem 3.2.1.

3.2.2 General Random Vectors

In this section, a more general case without normality assumption is considered.

Assumption 3.2.1. Let $\{\mathbf{V}_n\}_{n=1}^N$ be a sequence of independent random vectors in \mathbb{R}^D , such that for $n = 1, 2, \dots, N$,

(a). The expected mean of the random vector is provided as follows,

$$\mathbb{E}[\mathbf{V}_n] = \alpha,$$

(b). The expected covariance matrix of the random vector is provided as follows,

$$\mathbb{E}[(\mathbf{V}_n - \alpha)(\mathbf{V}_n - \alpha)^\top] = \Omega_D.$$

The expected mean of the random vector is known

We first consider that we know the exact value of the expected mean. If α is known, vectors can be easily centralized. Define the normalized sum with centralization as

$$S_{N,\alpha}^V := \frac{1}{\sqrt{N}} \sum_{n=1}^N (\mathbf{V}_n - \alpha).$$

The high dimensional central limit theorem from Chernozhukov et al. (2017) will be applied, since the dimension D is no longer constant, but tends to approach infinity. As a result of it, the following assumptions are required.

Assumption 3.2.2. *There are constant b , a sequence of constants $B_N \geq 1$ and a covariance estimator $\hat{\Omega}_{N,D}$ for $\{\mathbf{V}_n - \alpha\}_{n=1}^N$, which satisfy the following conditions:*

(a).

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E}[(\mathbf{V}_{n,d} - \alpha_d)^2] \geq b \text{ for all } d = 1, 2, \dots, D,$$

(b).

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E}[|\mathbf{V}_{n,d} - \alpha_d|^{2+k}] \leq B_N^k \text{ for all } d = 1, 2, \dots, D \text{ and } k = 1, 2,$$

(c).

$$\mathbb{E}[\exp(|\mathbf{V}_{n,d} - \alpha_d|/B_N)] \leq 2 \text{ for all } d = 1, 2, \dots, D \text{ and } n = 1, 2, \dots, N,$$

(d).

$$\left(\frac{B_N^2 \log^7(DN)}{N} \right) = o_p(1),$$

Then, we proposed out test statistics as the following:

$$T_3 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \hat{\Omega}_{N,D}^{-\frac{1}{2}} S_{N,\alpha}^V \right\|^2 - 1 \right). \quad (3.2.3)$$

Theorem 3.2.3. *Based on the central limit theorem from Chernozhukov et al. (2017), and Assumptions in 3.2.1 and 3.2.2, we have*

$$T_3 \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } N, D \rightarrow \infty.$$

Proof: Suppose $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_N$ are independent centered normal random vectors in \mathcal{R}^D , such that each \mathbf{W}_n has the same covariance matrix as \mathbf{V}_n , that is,

$$\mathbf{W}_n \sim \mathcal{N}(0, \Omega_D),$$

with the normalized sum $S_N^W := \frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{W}_n$. Under the Assumptions 3.2.2 a to d, it can be shown that $S_{N,\alpha}^V$ converges to S_N^W in distribution. As a result of $\Omega_D^{-\frac{1}{2}} \mathbf{W}_n \sim \mathcal{N}(0, I_{D \times D})$, Theorem 3.2.2 gives us

$$\sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^N \Omega_D^{-\frac{1}{2}} \mathbf{W}_n \right\|^2 - 1 \right) \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } N, D \rightarrow \infty.$$

Based on the continuous mapping theorem and Assumption 3.2.2 d, there is the result in Theorem 3.2.3. \square

The expected mean of the random vector is unknown

Then we consider that we do not know the exact value of the expected mean. Define the normalized sum as

$$S_N^V := \frac{1}{\sqrt{N}} \sum_{n=1}^N (\mathbf{V}_n - \bar{\mathbf{V}}_N),$$

where $\bar{\mathbf{V}}_N = \frac{1}{N} \sum_{n=1}^N \mathbf{V}_n$.

Then, we proposed our test statistics as the following:

$$T_4 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \hat{\Omega}_{N,D}^{-\frac{1}{2}} S_N^V \right\|^2 - 1 \right), \quad (3.2.4)$$

where $\hat{\Omega}_{N,D}$ is the covariance estimator for $\{\mathbf{V}_n - \bar{\mathbf{V}}_N\}_{n=1}^N$.

Theorem 3.2.4. *Based on the central limit theorem from Chernozhukov et al. (2017), Assumptions in 3.2.1 and 3.2.2, and Theorem 3.2.3, we have*

$$T_4 \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } N, D \rightarrow \infty.$$

As illustrates by Chernozhukov et al. (2017), S_N^V converges in distribution to $S_{N,\alpha}^V := \frac{1}{\sqrt{N}} \sum_{n=1}^N (\mathbf{V}_n - \alpha)$ as $N, D \rightarrow \infty$. Therefore, the result in Theorem 3.2.4 is straightforward.

It is worth noting that there are four different test statistics introduced in this section for different cases of random vectors respectively, but they are foundationally similar. The only difference is the procedure of standardization under different scenarios.

3.3 Implementation

In this section, we introduce how to apply our proposed test in bootstrap.

3.3.1 Multivariate normal random vectors with identity covariance matrix

First, we focus on the scenario that the random vectors are i.i.d. multivariate normal random vectors with identity covariance matrix. According to the test statistics in section 3.2.1, suppose one observes a realization of i.i.d. multivariate normal random vectors with an identity covariance matrix $\{\mathbf{v}_n\}_{n=1}^N$ of $\{\mathbf{V}_n\}_{n=1}^N$ and aims to check whether this data is under high dimensional conditions or not. We propose our test with multiplier bootstrap. The steps are listed as follows,

- (i) Generate an sample e_1, e_2, \dots, e_N from $\mathcal{N}(0, 1)$, which is independent to \mathbf{v}_n ,
- (ii) Consider a new list of random vectors as $\mathbf{v}_{n,B} = e_n \mathbf{v}_n$,
- (iii) Calculate the test statistic $T_1 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \mathbf{v}_{n,B} \right) \right\|^2 - 1 \right)$,
- (iv) Repeat (i) to (iii) for B times,
- (v) Use a Kolmogorov–Smirnov (KS) test to examine whether the distribution of the test statistics from the bootstrap is a distribution of $\mathcal{N}(0, 1)$.

If we reject the null hypothesis that the distribution of the test statistics is a standard normal, then we reject the null hypothesis that random vectors is in high dimensionality.

3.3.2 Multivariate normal random vectors with non-identity covariance matrix

Then, we focus on the scenario that the random vectors are multivariate normal random vectors, not with an identity covariance matrix. According to the test statistics in section 3.2.2, suppose one observes a realization of i.i.d. multivariate normal random vectors $\{\mathbf{v}_n\}_{n=1}^N$ of $\{\mathbf{V}_n\}_{n=1}^N$ and aims to check whether this data is under high dimensional conditions or not. We propose our test with multiplier bootstrap. The steps are listed as

follows,

- (i) Calculate the estimates of covariance matrix $\hat{\Omega}$ from $\{\mathbf{v}_n\}_{n=1}^N$,
- (ii) Generate an sample e_1, e_2, \dots, e_N from $\mathcal{N}(0, 1)$, which is independent to \mathbf{v}_n ,
- (iii) Consider a new list of random vectors as $\mathbf{v}_{n,B} = e_n \mathbf{v}_n$,
- (iv) Calculate the test statistic $T_2 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \hat{\Omega}^{-\frac{1}{2}} \mathbf{v}_{n,B} \right) \right\|^2 - 1 \right)$,
- (v) Repeat (ii) to (iv) for B times,
- (vi) Use a KS test to examine whether the distribution of the test statistics from the bootstrap is a distribution of $\mathcal{N}(0, 1)$.

If we reject the null hypothesis that the distribution of the test statistics is a standard normal, then we reject the null hypothesis that random vectors is in high dimensionality.

3.3.3 General random vectors with known expected mean

Now, we focus on the general random vectors with known expected mean. According to the test statistics in section 3.2.3, suppose one observes a realization of i.i.d. multivariate normal random vectors $\{\mathbf{v}_n\}_{n=1}^N$ of $\{\mathbf{V}_n\}_{n=1}^N$ and aims to check whether this data is under high dimensional conditions or not. We propose out test with multiplier bootstrap. The steps are listed as follows,

- (i) Calculate the estimates of covariance matrix $\hat{\Omega}$ from $\{\mathbf{v}_n - \alpha\}_{n=1}^N$,
- (ii) Generate an sample e_1, e_2, \dots, e_N from $\mathcal{N}(0, 1)$, which is independent to \mathbf{v}_n ,
- (iii) Consider new random vectors as $\mathbf{v}_{n,B} = e_n(\mathbf{v}_n - \alpha)$, where α is the known expected mean,
- (iv) Calculate the test statistic $t_3 = \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \hat{\Omega}^{-\frac{1}{2}} \frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{v}_{n,B} \right\|^2 - 1 \right)$,
- (v) Repeat (ii) to (iv) for B times,
- (vi) Use a KS test to examine whether the distribution of the test statistics from the bootstrap is a distribution of $\mathcal{N}(0, 1)$.

If we reject the null hypothesis that the distribution of the test statistics is a standard normal, then we reject the null hypothesis that random vectors is in high dimensionality.

3.3.4 General random vectors with unknown expected mean

Finally, we focus on the General random vectors with unknown expected mean. According to the test statistics in section 3.2.4, suppose one observes a realization of i.i.d. multivariate normal random vectors $\{\mathbf{v}_n\}_{n=1}^N$ of $\{\mathbf{V}_n\}_{n=1}^N$ and aims to check whether this data is under high dimensional conditions or not. We propose out test with multiplier bootstrap. The steps are listed as follows,

- (i) Calculate the estimated mean $\bar{\mathbf{v}}_N$ from $\{\mathbf{v}_n\}_{n=1}^N$,
- (ii) Calculate the estimated covariance matrix $\hat{\Omega}$ from $\{\mathbf{v}_n - \bar{\mathbf{v}}_N\}_{n=1}^N$,
- (iii) Generate an sample e_1, e_2, \dots, e_N from $\mathcal{N}(0, 1)$, which is independent to \mathbf{v}_n ,
- (iv) Consider a new list of random vectors as $\mathbf{v}_{n,B} = e_n \mathbf{v}_n$,
- (v) Calculate the test statistic $t_4 = \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \hat{\Omega}^{-\frac{1}{2}} \frac{1}{\sqrt{N}} \sum_{n=1}^N (\mathbf{v}_{n,B} - \bar{\mathbf{v}}_N) \right\|^2 - 1 \right)$,
- (vi) Repeat (iii) to (v) for B times,
- (vii) Use a KS test to examine whether the distribution of the test statistics from the bootstrap is a distribution of $\mathcal{N}(0, 1)$.

If we reject the null hypothesis that the distribution of the test statistics is a standard normal, then we reject the null hypothesis that random vectors is in high dimensionality.

3.4 Monte Carlo Study

In this section, we present a Monte Carlo study to illustrate the size and power of our test. We consider different designs for the tests of random vectors. We set different sample sizes N , from 100 to 2000. Within each case, the dimension D is chosen from 2 to 200. To keep the running time manageable, these results are based on 1000 simulations. Within each simulation, we set the bootstrap times $B = 500$.

3.4.1 Simulation Designs

Suppose there is a sequence of D -dimensional random vectors $\{\mathbf{V}_n\}_{n=1}^N$. We proposed five different designs for this simulation study.

For **Design (i)**, we propose our HD test on the i.i.d. multivariate normal random vectors with identity covariance matrix, such that

$$\mathbf{V}_n \sim \mathcal{N}(0, I_{D \times D}).$$

In each simulation, with the estimated mean $\bar{\mathbf{V}}_N$ and estimated covariance matrix $\hat{\Omega}_D$ for the sample under high dimension conditions, we calculate the sampling distribution of our test statistics with bootstrap according to Section 3.3.4, and test whether the distribution is a standard normal using KS test at 5% level. Then we calculate the overall rejection rate among 1000 simulations.

For **Design (ii)**, we propose our HD test on the i.i.d. multivariate normal random vectors with non-identity covariance matrix, such that

$$\mathbf{V}_n \sim \mathcal{N}(0, \Omega_D),$$

where Ω_D is a matrix in which the entries outside the main diagonal are all zero,

$$\Omega_{D,ii} = i, \text{ for } i = 1, 2, \dots, D.$$

Here, $\Omega_{D,ii}$ is the element in i^{th} row and i^{th} column of the matrix Ω_D . In each simulation, with the estimated mean $\bar{\mathbf{V}}_N$ and estimated covariance matrix $\hat{\Omega}_D$ for the sample under high dimension conditions, we calculate the sampling distribution of our test statistics with bootstrap according to Section 3.3.4, and test whether the distribution is a standard normal using KS test at 5% level. Then we calculate the overall rejection rate among 1000 simulations.

For **Design (iii)**, we propose our HD test on the general random vectors, such that

$$\begin{aligned} \mathbf{V}_n &\sim \mathcal{N}(\alpha, \Omega_D), \\ \alpha &= (1, 2, \dots, D)'. \end{aligned}$$

Ω_D is a matrix in which the entries outside the main diagonal are all zero,

$$\Omega_{D,ii} = i, \text{ for } i = 1, 2, \dots, D,$$

where $\Omega_{D,ii}$ is the element in i^{th} row and i^{th} column of the matrix Ω_D . In each simulation, with the estimated mean $\bar{\mathbf{V}}_N$ and estimated covariance matrix $\hat{\Omega}_D$ for the sample under high dimension conditions, we calculate the sampling distribution of our test statistics with bootstrap according to Section 3.3.4, and test whether the distribution is a standard normal using KS test at 5% level. Then we calculate the overall rejection rate among 1000 simulations.

For **Design (iv)**, we propose our HD test on the general random vectors, such that

$$\begin{aligned} \mathbf{V}_n &\sim \mathcal{N}(\alpha, \Omega_D), \\ \alpha &= (1, 2, \dots, D)' \end{aligned}$$

Ω_D is a matrix such as,

$$\begin{aligned} \Omega_{D,ii} &= 1, \text{ for } i = 1, 2, \dots, D, \\ \Omega_{D,i(i+1)} &= \Omega_{D,(i+1)i} = 0.2, \text{ for } i = 1, 2, \dots, D - 1, \end{aligned}$$

where $\Omega_{D,ij}$ is the element in i^{th} row and j^{th} column of the matrix Ω_D , and $i, j = 1, 2, \dots, D$. The remaining entries of Ω_D are all zero. In each simulation, with the estimated mean $\bar{\mathbf{V}}_N$ and estimated covariance matrix $\hat{\Omega}_D$ for the sample under high dimension conditions, we calculate the sampling distribution of our test statistics with bootstrap according to Section 3.3.4, and test whether the distribution is a standard normal using KS test at 5% level. Then we calculate the overall rejection rate among 1000 simulations.

For **Design (v)**, we propose our HD test on a more general random vectors, such that

$$\begin{aligned}
\mathbb{E}[\mathbf{V}_n] &= \alpha, \text{ for } \alpha = (1, 2, \dots, D)', \\
\mathbb{E}[(\mathbf{V}_n - \alpha)(\mathbf{V}_n - \alpha)^\top] &= \Omega_D, \\
\mathbf{V}_n^{(i)} &\sim \Gamma(i, 1), \text{ for } i = 1, 2, \dots, \lfloor \frac{D}{2} \rfloor, \\
\mathbf{V}_n^{(i)} &\sim \mathbf{U}(0.5i, i), \text{ for } i = \lfloor \frac{D}{2} \rfloor + 1, \lfloor \frac{D}{2} \rfloor + 2, \dots, D, \\
\text{corr}(\mathbf{V}_n^{(i)}, \mathbf{V}_n^{(i+1)}) &= 0.2, \text{ for } i = 1, 2, \dots, D - 1, \\
\text{corr}(\mathbf{V}_n^{(i)}, \mathbf{V}_n^{(j)}) &= 0, \text{ for } |i - j| > 1,
\end{aligned}$$

where $\mathbf{V}_n^{(i)}$ is the i^{th} element in the vector \mathbf{V}_n and $i = 1, 2, \dots, D$. In each simulation, with the estimated mean $\bar{\mathbf{V}}_N$ and estimated covariance matrix $\hat{\Omega}_D$ for the sample under high dimension conditions, we calculate the sampling distribution of our test statistics with bootstrap according to Section 3.3.4, and test whether the distribution is a standard normal using KS test at 5% level. Then we calculate the overall rejection rate among 1000 simulations.

3.4.2 Results and discussion

In each round of simulation, $B = 500$ times of bootstrap generates 500 test statistics. KS test is applied to test whether the empirical sampling distribution of our proposed test statistics is a standard normal. Among the 1000 rounds of simulation, the rejection rate of rejecting the null hypothesis at 5% in KS test is calculated. Theoretically, the rejection rate is very close to 5% if the asymptotic distribution is a standard normal, and 5% is also the threshold to reject the null hypothesis of our proposed HD test as the asymptotic distribution of our proposed test statistics is a standard normal under high dimension conditions. However, the rejection rate itself, generated from a bootstrap procedure with sufficient times, represents the similarity between the empirical sampling distribution and the underlying distribution we test. Such similarity depends on how the rejection rate is close to 5% (e.g. Horowitz, 2019; Davidson and MacKinnon, 2006). Therefore, the threshold to reject the null hypothesis of our HD test is not necessarily to be 5%, and how the rejection rate close to 5% means how the sample data is close to the high dimensionality. In this section, we consider two thresholds, 5% and 10%.

Tables 3.1 illustrates: (1) the empirical rejection rates approach 1 as D decreases to 2; (2) the empirical rejection rates do not decrease to 0.05 when N is less than or equal to 200 and even D increases to N ; (3) given a sample size N , the rejection rate decreases as D increases; (4) given a sample size N , the decline speed of the rejection rate defers among different N , and rejection rate has a faster decreasing speed with a larger sample size N ; (5) the empirical rejection rates decreased from 1 to approximately 0.05 when N and D both becomes larger enough; (6) if the sample size N is large enough, it appears that how large D need to be for high dimensionality is independent to the sample size N . Such conclusions can also be found in Table 3.2, 3.3, 3.4 and 3.5.

Given the results of (6), we define D^* as the “dividing line” of threshold 10% for the empirical rejection rates under the given sample size N , such that the rejection rate for almost all D that small than D^* is equal or larger than 10%, and the rejection rate for almost all D that equal or larger than D^* is smaller than 10%. We also define D^{**} as the “dividing line” of threshold 5% for the empirical rejection rates under the given sample size N , such that the rejection rate for almost all D that small than D^* is equal or larger than 5%, and the rejection rate for almost all D that equal or larger than D^* is smaller than 5%. D^* represents that, given a sample size N , the smallest D we can not reject the null hypothesis of high dimensionality of random vectors at 10%. D^{**} represents that, given a sample size N , the smallest D we can not reject the null hypothesis of high dimensionality of random vectors at 5%. The trends of the “dividing line” D^* and D^{**} with the increasing sample size N for five designs are demonstrated in the Table 3.6, 3.7, 3.8, 3.9 and 3.10.

The results reveal that: (1) D^* show a apparent decreasing trend (if applicable) as the sample size N increases; (2) D^{**} also show a apparent decreasing trend (if applicable) as the sample size N increases; (3) the decline rates of both D^* and D^{**} are becoming slight as the sample size N increases; (4) Both D^* and D^{**} appear to converge to a specific value as the sample size N is large enough, which is correlated with the underlying distribution of the vector \mathbf{V}_n ; (5) As the underlying distribution of the vector \mathbf{V}_n becomes more and more complicated from Design (i) to Design (v), the empirical converged values of D^* and D^{**} are increased.

Results (1)(2) illustrate that under a relatively small sample size N , the value of D^* and D^{**} is highly correlated with the sample size. Both D^* and D^{**} (if applicable) decrease with the growing sample size with some random fluctuations. Results (3)(4) show that under a

relatively large sample size N , the value of D^* and D^{**} is no longer highly correlated with the sample size, but rather the underlying distribution of the vector \mathbf{V}_n . Intuitively, if we observed enough number of random vectors, only the underlying distribution of the random vectors in the finite sample alters the dimension that is needed for high dimensionality. If the underlying distribution is more complex, a larger dimension of random vectors is required for high dimensionality as the cost to achieve the high dimension properties.

Table 3.1: Empirical rejection rates in Designs (i)

| D | Sample size N | | | | | | | | | | | |
|-----|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| 2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 0.628 | 0.592 | 0.566 | 0.596 | 0.538 | 0.610 | 0.586 | 0.566 | 0.588 | 0.568 | 0.590 | 0.558 |
| 4 | 0.476 | 0.472 | 0.454 | 0.462 | 0.474 | 0.468 | 0.466 | 0.458 | 0.450 | 0.406 | 0.450 | 0.502 |
| 5 | 0.468 | 0.384 | 0.416 | 0.352 | 0.358 | 0.346 | 0.368 | 0.376 | 0.358 | 0.366 | 0.444 | 0.376 |
| 6 | 0.400 | 0.366 | 0.338 | 0.370 | 0.372 | 0.338 | 0.340 | 0.306 | 0.354 | 0.376 | 0.326 | 0.320 |
| 7 | 0.364 | 0.308 | 0.300 | 0.246 | 0.292 | 0.302 | 0.306 | 0.304 | 0.270 | 0.300 | 0.260 | 0.278 |
| 8 | 0.302 | 0.274 | 0.278 | 0.280 | 0.296 | 0.242 | 0.284 | 0.232 | 0.232 | 0.262 | 0.252 | 0.296 |
| 10 | 0.336 | 0.278 | 0.234 | 0.252 | 0.236 | 0.264 | 0.240 | 0.240 | 0.226 | 0.272 | 0.252 | 0.232 |
| 20 | 0.276 | 0.254 | 0.210 | 0.212 | 0.216 | 0.214 | 0.208 | 0.228 | 0.196 | 0.184 | 0.184 | 0.218 |
| 30 | 0.216 | 0.174 | 0.166 | 0.138 | 0.154 | 0.146 | 0.132 | 0.142 | 0.124 | 0.130 | 0.100 | 0.114 |
| 40 | 0.184 | 0.110 | 0.138 | 0.098 | 0.130 | 0.122 | 0.104 | 0.108 | 0.090 | 0.096 | 0.098 | 0.092 |
| 50 | 0.164 | 0.126 | 0.120 | 0.112 | 0.104 | 0.120 | 0.106 | 0.098 | 0.086 | 0.092 | 0.092 | 0.082 |
| 60 | 0.174 | 0.098 | 0.132 | 0.092 | 0.080 | 0.068 | 0.106 | 0.108 | 0.072 | 0.120 | 0.068 | 0.080 |
| 70 | 0.188 | 0.150 | 0.116 | 0.086 | 0.090 | 0.104 | 0.100 | 0.098 | 0.112 | 0.070 | 0.076 | 0.098 |
| 80 | 0.150 | 0.128 | 0.104 | 0.092 | 0.094 | 0.104 | 0.080 | 0.098 | 0.084 | 0.072 | 0.088 | 0.072 |
| 90 | 0.154 | 0.110 | 0.094 | 0.104 | 0.088 | 0.070 | 0.074 | 0.068 | 0.082 | 0.090 | 0.066 | 0.060 |
| 100 | 0.142 | 0.120 | 0.102 | 0.078 | 0.092 | 0.090 | 0.066 | 0.066 | 0.078 | 0.074 | 0.078 | 0.064 |
| 110 | - | 0.100 | 0.108 | 0.092 | 0.078 | 0.062 | 0.066 | 0.076 | 0.080 | 0.058 | 0.072 | 0.052 |
| 120 | - | 0.136 | 0.100 | 0.078 | 0.086 | 0.062 | 0.068 | 0.074 | 0.072 | 0.074 | 0.069 | 0.068 |
| 130 | - | 0.140 | 0.108 | 0.082 | 0.100 | 0.092 | 0.064 | 0.082 | 0.068 | 0.082 | 0.064 | 0.061 |
| 140 | - | 0.098 | 0.102 | 0.072 | 0.078 | 0.078 | 0.072 | 0.062 | 0.063 | 0.062 | 0.072 | 0.065 |
| 150 | - | 0.102 | 0.094 | 0.080 | 0.098 | 0.064 | 0.066 | 0.062 | 0.059 | 0.062 | 0.054 | 0.058 |
| 160 | - | 0.092 | 0.078 | 0.106 | 0.074 | 0.074 | 0.068 | 0.076 | 0.052 | 0.076 | 0.048 | 0.049 |
| 170 | - | 0.108 | 0.082 | 0.080 | 0.074 | 0.068 | 0.082 | 0.054 | 0.063 | 0.062 | 0.048 | 0.042 |
| 180 | - | 0.098 | 0.086 | 0.070 | 0.082 | 0.096 | 0.088 | 0.058 | 0.049 | 0.068 | 0.047 | 0.041 |
| 190 | - | 0.110 | 0.092 | 0.080 | 0.072 | 0.074 | 0.066 | 0.060 | 0.052 | 0.049 | 0.050 | 0.049 |
| 200 | - | 0.126 | 0.106 | 0.076 | 0.092 | 0.066 | 0.068 | 0.046 | 0.049 | 0.047 | 0.046 | 0.048 |

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05.

Table 3.2: Empirical rejection rates in Designs (ii)

| D | Sample size N | | | | | | | | | | | |
|-----|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| 2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 0.671 | 0.585 | 0.608 | 0.593 | 0.572 | 0.632 | 0.625 | 0.561 | 0.634 | 0.576 | 0.589 | 0.592 |
| 4 | 0.497 | 0.469 | 0.447 | 0.463 | 0.499 | 0.493 | 0.489 | 0.473 | 0.450 | 0.410 | 0.472 | 0.541 |
| 5 | 0.477 | 0.412 | 0.418 | 0.362 | 0.358 | 0.355 | 0.397 | 0.377 | 0.383 | 0.376 | 0.474 | 0.388 |
| 6 | 0.399 | 0.364 | 0.336 | 0.398 | 0.393 | 0.362 | 0.336 | 0.301 | 0.358 | 0.373 | 0.341 | 0.329 |
| 7 | 0.360 | 0.327 | 0.316 | 0.254 | 0.306 | 0.302 | 0.320 | 0.321 | 0.291 | 0.312 | 0.273 | 0.278 |
| 8 | 0.300 | 0.272 | 0.287 | 0.296 | 0.298 | 0.261 | 0.290 | 0.240 | 0.239 | 0.279 | 0.271 | 0.313 |
| 10 | 0.362 | 0.287 | 0.234 | 0.267 | 0.240 | 0.259 | 0.255 | 0.237 | 0.227 | 0.282 | 0.260 | 0.244 |
| 20 | 0.292 | 0.265 | 0.213 | 0.227 | 0.219 | 0.217 | 0.224 | 0.225 | 0.203 | 0.191 | 0.187 | 0.216 |
| 30 | 0.222 | 0.180 | 0.172 | 0.139 | 0.153 | 0.152 | 0.134 | 0.149 | 0.133 | 0.140 | 0.108 | 0.123 |
| 40 | 0.198 | 0.118 | 0.136 | 0.098 | 0.139 | 0.128 | 0.103 | 0.108 | 0.089 | 0.100 | 0.105 | 0.114 |
| 50 | 0.175 | 0.134 | 0.125 | 0.118 | 0.103 | 0.124 | 0.112 | 0.105 | 0.089 | 0.091 | 0.123 | 0.085 |
| 60 | 0.179 | 0.106 | 0.132 | 0.095 | 0.086 | 0.070 | 0.107 | 0.110 | 0.072 | 0.129 | 0.071 | 0.086 |
| 70 | 0.185 | 0.161 | 0.121 | 0.086 | 0.095 | 0.112 | 0.100 | 0.100 | 0.116 | 0.073 | 0.080 | 0.099 |
| 80 | 0.161 | 0.133 | 0.111 | 0.096 | 0.101 | 0.107 | 0.086 | 0.097 | 0.086 | 0.073 | 0.091 | 0.076 |
| 90 | 0.163 | 0.109 | 0.098 | 0.107 | 0.091 | 0.070 | 0.076 | 0.068 | 0.087 | 0.108 | 0.069 | 0.059 |
| 100 | 0.144 | 0.126 | 0.105 | 0.081 | 0.095 | 0.096 | 0.071 | 0.066 | 0.083 | 0.079 | 0.083 | 0.063 |
| 110 | - | 0.106 | 0.113 | 0.092 | 0.077 | 0.063 | 0.070 | 0.080 | 0.082 | 0.062 | 0.076 | 0.055 |
| 120 | - | 0.137 | 0.099 | 0.083 | 0.086 | 0.064 | 0.072 | 0.077 | 0.076 | 0.078 | 0.074 | 0.069 |
| 130 | - | 0.149 | 0.112 | 0.082 | 0.105 | 0.098 | 0.068 | 0.083 | 0.067 | 0.088 | 0.068 | 0.065 |
| 140 | - | 0.098 | 0.104 | 0.073 | 0.081 | 0.081 | 0.077 | 0.065 | 0.065 | 0.067 | 0.072 | 0.069 |
| 150 | - | 0.107 | 0.093 | 0.079 | 0.105 | 0.065 | 0.068 | 0.062 | 0.064 | 0.061 | 0.056 | 0.060 |
| 160 | - | 0.092 | 0.082 | 0.110 | 0.077 | 0.073 | 0.070 | 0.076 | 0.056 | 0.080 | 0.049 | 0.049 |
| 170 | - | 0.107 | 0.086 | 0.081 | 0.075 | 0.071 | 0.081 | 0.058 | 0.066 | 0.062 | 0.045 | 0.044 |
| 180 | - | 0.105 | 0.122 | 0.071 | 0.085 | 0.097 | 0.095 | 0.058 | 0.051 | 0.068 | 0.048 | 0.044 |
| 190 | - | 0.111 | 0.092 | 0.080 | 0.074 | 0.073 | 0.070 | 0.062 | 0.055 | 0.045 | 0.048 | 0.049 |
| 200 | - | 0.086 | 0.093 | 0.078 | 0.096 | 0.070 | 0.067 | 0.049 | 0.049 | 0.050 | 0.049 | 0.047 |

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05.

Table 3.3: Empirical rejection rates in Designs (iii)

| D | Sample size N | | | | | | | | | | | |
|-----|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| 2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 0.659 | 0.611 | 0.544 | 0.617 | 0.521 | 0.626 | 0.619 | 0.560 | 0.619 | 0.553 | 0.620 | 0.579 |
| 4 | 0.481 | 0.467 | 0.440 | 0.457 | 0.476 | 0.462 | 0.449 | 0.464 | 0.466 | 0.420 | 0.445 | 0.483 |
| 5 | 0.490 | 0.382 | 0.440 | 0.344 | 0.362 | 0.355 | 0.381 | 0.367 | 0.357 | 0.352 | 0.439 | 0.388 |
| 6 | 0.399 | 0.371 | 0.334 | 0.363 | 0.390 | 0.339 | 0.338 | 0.316 | 0.350 | 0.397 | 0.319 | 0.322 |
| 7 | 0.350 | 0.317 | 0.295 | 0.246 | 0.281 | 0.294 | 0.299 | 0.301 | 0.282 | 0.311 | 0.253 | 0.274 |
| 8 | 0.298 | 0.270 | 0.283 | 0.287 | 0.295 | 0.247 | 0.274 | 0.230 | 0.224 | 0.271 | 0.245 | 0.301 |
| 10 | 0.339 | 0.279 | 0.237 | 0.244 | 0.246 | 0.255 | 0.234 | 0.246 | 0.226 | 0.264 | 0.257 | 0.225 |
| 20 | 0.281 | 0.250 | 0.219 | 0.222 | 0.219 | 0.216 | 0.218 | 0.233 | 0.204 | 0.194 | 0.187 | 0.227 |
| 30 | 0.211 | 0.179 | 0.161 | 0.142 | 0.148 | 0.141 | 0.137 | 0.142 | 0.125 | 0.126 | 0.100 | 0.111 |
| 40 | 0.188 | 0.109 | 0.144 | 0.098 | 0.130 | 0.120 | 0.101 | 0.109 | 0.089 | 0.099 | 0.098 | 0.094 |
| 50 | 0.172 | 0.121 | 0.123 | 0.111 | 0.101 | 0.126 | 0.110 | 0.097 | 0.087 | 0.093 | 0.091 | 0.125 |
| 60 | 0.182 | 0.094 | 0.128 | 0.094 | 0.084 | 0.071 | 0.105 | 0.107 | 0.071 | 0.121 | 0.115 | 0.085 |
| 70 | 0.181 | 0.154 | 0.123 | 0.091 | 0.095 | 0.110 | 0.099 | 0.097 | 0.109 | 0.069 | 0.079 | 0.094 |
| 80 | 0.147 | 0.124 | 0.105 | 0.095 | 0.098 | 0.105 | 0.077 | 0.102 | 0.089 | 0.070 | 0.093 | 0.074 |
| 90 | 0.153 | 0.114 | 0.092 | 0.104 | 0.091 | 0.072 | 0.078 | 0.070 | 0.086 | 0.114 | 0.064 | 0.060 |
| 100 | 0.150 | 0.123 | 0.108 | 0.076 | 0.093 | 0.092 | 0.067 | 0.070 | 0.080 | 0.073 | 0.075 | 0.064 |
| 110 | - | 0.097 | 0.109 | 0.093 | 0.079 | 0.064 | 0.065 | 0.075 | 0.077 | 0.061 | 0.073 | 0.052 |
| 120 | - | 0.140 | 0.105 | 0.080 | 0.087 | 0.065 | 0.067 | 0.072 | 0.072 | 0.073 | 0.067 | 0.067 |
| 130 | - | 0.138 | 0.113 | 0.080 | 0.102 | 0.093 | 0.067 | 0.086 | 0.071 | 0.087 | 0.068 | 0.062 |
| 140 | - | 0.100 | 0.104 | 0.069 | 0.077 | 0.079 | 0.071 | 0.064 | 0.063 | 0.063 | 0.072 | 0.065 |
| 150 | - | 0.102 | 0.097 | 0.078 | 0.104 | 0.066 | 0.069 | 0.065 | 0.061 | 0.061 | 0.053 | 0.059 |
| 160 | - | 0.093 | 0.076 | 0.111 | 0.078 | 0.071 | 0.071 | 0.076 | 0.051 | 0.079 | 0.054 | 0.052 |
| 170 | - | 0.109 | 0.084 | 0.079 | 0.074 | 0.070 | 0.080 | 0.054 | 0.066 | 0.061 | 0.048 | 0.042 |
| 180 | - | 0.103 | 0.085 | 0.073 | 0.079 | 0.098 | 0.093 | 0.060 | 0.049 | 0.070 | 0.046 | 0.041 |
| 190 | - | 0.107 | 0.097 | 0.077 | 0.073 | 0.076 | 0.065 | 0.060 | 0.051 | 0.052 | 0.049 | 0.048 |
| 200 | - | 0.085 | 0.109 | 0.074 | 0.094 | 0.066 | 0.066 | 0.047 | 0.050 | 0.056 | 0.045 | 0.049 |

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05.

Table 3.4: Empirical rejection rates in Designs (iv)

| D | Sample size N | | | | | | | | | | | |
|-----|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| 2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 0.688 | 0.613 | 0.610 | 0.624 | 0.604 | 0.682 | 0.609 | 0.637 | 0.627 | 0.635 | 0.622 | 0.608 |
| 4 | 0.503 | 0.490 | 0.477 | 0.480 | 0.503 | 0.494 | 0.484 | 0.487 | 0.481 | 0.450 | 0.464 | 0.537 |
| 5 | 0.505 | 0.396 | 0.447 | 0.392 | 0.394 | 0.365 | 0.389 | 0.416 | 0.392 | 0.410 | 0.463 | 0.424 |
| 6 | 0.429 | 0.392 | 0.365 | 0.408 | 0.420 | 0.375 | 0.367 | 0.342 | 0.383 | 0.410 | 0.352 | 0.348 |
| 7 | 0.381 | 0.348 | 0.320 | 0.275 | 0.330 | 0.328 | 0.319 | 0.334 | 0.286 | 0.315 | 0.286 | 0.300 |
| 8 | 0.313 | 0.305 | 0.307 | 0.308 | 0.326 | 0.253 | 0.318 | 0.244 | 0.244 | 0.279 | 0.273 | 0.328 |
| 10 | 0.367 | 0.301 | 0.252 | 0.265 | 0.251 | 0.293 | 0.252 | 0.259 | 0.240 | 0.293 | 0.264 | 0.246 |
| 20 | 0.301 | 0.278 | 0.226 | 0.231 | 0.229 | 0.221 | 0.227 | 0.252 | 0.211 | 0.190 | 0.200 | 0.225 |
| 30 | 0.237 | 0.182 | 0.171 | 0.154 | 0.169 | 0.156 | 0.144 | 0.149 | 0.135 | 0.141 | 0.110 | 0.119 |
| 40 | 0.190 | 0.114 | 0.155 | 0.101 | 0.140 | 0.131 | 0.109 | 0.112 | 0.101 | 0.106 | 0.103 | 0.129 |
| 50 | 0.182 | 0.135 | 0.125 | 0.121 | 0.114 | 0.124 | 0.117 | 0.106 | 0.094 | 0.103 | 0.107 | 0.111 |
| 60 | 0.181 | 0.101 | 0.140 | 0.098 | 0.084 | 0.075 | 0.118 | 0.120 | 0.077 | 0.133 | 0.111 | 0.128 |
| 70 | 0.204 | 0.157 | 0.123 | 0.092 | 0.094 | 0.111 | 0.109 | 0.104 | 0.125 | 0.135 | 0.109 | 0.102 |
| 80 | 0.160 | 0.135 | 0.112 | 0.098 | 0.099 | 0.117 | 0.083 | 0.108 | 0.091 | 0.079 | 0.095 | 0.079 |
| 90 | 0.172 | 0.121 | 0.097 | 0.114 | 0.098 | 0.075 | 0.078 | 0.075 | 0.087 | 0.099 | 0.072 | 0.064 |
| 100 | 0.150 | 0.131 | 0.106 | 0.088 | 0.101 | 0.098 | 0.072 | 0.069 | 0.085 | 0.081 | 0.087 | 0.071 |
| 110 | - | 0.109 | 0.120 | 0.099 | 0.086 | 0.064 | 0.068 | 0.082 | 0.085 | 0.063 | 0.076 | 0.057 |
| 120 | - | 0.146 | 0.112 | 0.081 | 0.097 | 0.065 | 0.072 | 0.080 | 0.077 | 0.083 | 0.074 | 0.076 |
| 130 | - | 0.145 | 0.119 | 0.092 | 0.111 | 0.100 | 0.071 | 0.091 | 0.073 | 0.085 | 0.067 | 0.066 |
| 140 | - | 0.102 | 0.112 | 0.077 | 0.086 | 0.086 | 0.078 | 0.070 | 0.066 | 0.067 | 0.074 | 0.072 |
| 150 | - | 0.106 | 0.104 | 0.083 | 0.110 | 0.067 | 0.070 | 0.065 | 0.064 | 0.064 | 0.060 | 0.060 |
| 160 | - | 0.100 | 0.084 | 0.115 | 0.082 | 0.083 | 0.071 | 0.082 | 0.056 | 0.080 | 0.053 | 0.055 |
| 170 | - | 0.117 | 0.091 | 0.083 | 0.083 | 0.072 | 0.088 | 0.059 | 0.068 | 0.068 | 0.058 | 0.054 |
| 180 | - | 0.109 | 0.094 | 0.073 | 0.085 | 0.100 | 0.099 | 0.060 | 0.051 | 0.070 | 0.049 | 0.043 |
| 190 | - | 0.123 | 0.101 | 0.083 | 0.080 | 0.077 | 0.069 | 0.062 | 0.054 | 0.053 | 0.047 | 0.047 |
| 200 | - | 0.092 | 0.114 | 0.085 | 0.096 | 0.074 | 0.074 | 0.050 | 0.056 | 0.050 | 0.049 | 0.047 |

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05.

Table 3.5: Empirical rejection rates in Designs (v)

| D | Sample size N | | | | | | | | | | | |
|-----|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| 2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 0.724 | 0.641 | 0.655 | 0.695 | 0.603 | 0.711 | 0.648 | 0.627 | 0.681 | 0.660 | 0.668 | 0.645 |
| 4 | 0.549 | 0.540 | 0.525 | 0.536 | 0.550 | 0.536 | 0.521 | 0.529 | 0.502 | 0.443 | 0.483 | 0.559 |
| 5 | 0.515 | 0.424 | 0.482 | 0.396 | 0.385 | 0.374 | 0.419 | 0.409 | 0.404 | 0.395 | 0.483 | 0.435 |
| 6 | 0.462 | 0.411 | 0.367 | 0.399 | 0.429 | 0.368 | 0.381 | 0.357 | 0.387 | 0.430 | 0.374 | 0.357 |
| 7 | 0.419 | 0.353 | 0.350 | 0.271 | 0.322 | 0.328 | 0.353 | 0.331 | 0.297 | 0.326 | 0.288 | 0.301 |
| 8 | 0.339 | 0.310 | 0.304 | 0.308 | 0.324 | 0.283 | 0.322 | 0.261 | 0.251 | 0.286 | 0.292 | 0.332 |
| 10 | 0.387 | 0.323 | 0.251 | 0.275 | 0.264 | 0.306 | 0.273 | 0.260 | 0.262 | 0.299 | 0.284 | 0.265 |
| 20 | 0.298 | 0.282 | 0.228 | 0.233 | 0.233 | 0.237 | 0.242 | 0.245 | 0.214 | 0.197 | 0.202 | 0.236 |
| 30 | 0.243 | 0.196 | 0.181 | 0.149 | 0.174 | 0.167 | 0.146 | 0.161 | 0.141 | 0.148 | 0.115 | 0.128 |
| 40 | 0.203 | 0.127 | 0.159 | 0.114 | 0.143 | 0.141 | 0.118 | 0.119 | 0.097 | 0.109 | 0.108 | 0.100 |
| 50 | 0.185 | 0.136 | 0.132 | 0.121 | 0.114 | 0.131 | 0.118 | 0.110 | 0.098 | 0.104 | 0.101 | 0.095 |
| 60 | 0.187 | 0.107 | 0.146 | 0.107 | 0.090 | 0.074 | 0.121 | 0.117 | 0.079 | 0.138 | 0.119 | 0.128 |
| 70 | 0.209 | 0.175 | 0.125 | 0.097 | 0.098 | 0.113 | 0.112 | 0.112 | 0.128 | 0.077 | 0.112 | 0.123 |
| 80 | 0.164 | 0.147 | 0.112 | 0.105 | 0.102 | 0.119 | 0.087 | 0.110 | 0.096 | 0.081 | 0.116 | 0.104 |
| 90 | 0.169 | 0.126 | 0.108 | 0.121 | 0.102 | 0.077 | 0.083 | 0.073 | 0.128 | 0.127 | 0.071 | 0.064 |
| 100 | 0.164 | 0.132 | 0.111 | 0.084 | 0.104 | 0.103 | 0.074 | 0.114 | 0.087 | 0.083 | 0.091 | 0.071 |
| 110 | - | 0.108 | 0.117 | 0.104 | 0.088 | 0.068 | 0.125 | 0.082 | 0.088 | 0.066 | 0.081 | 0.057 |
| 120 | - | 0.153 | 0.113 | 0.087 | 0.100 | 0.069 | 0.106 | 0.084 | 0.083 | 0.083 | 0.077 | 0.078 |
| 130 | - | 0.152 | 0.125 | 0.090 | 0.108 | 0.099 | 0.070 | 0.095 | 0.079 | 0.095 | 0.073 | 0.067 |
| 140 | - | 0.109 | 0.112 | 0.081 | 0.085 | 0.087 | 0.080 | 0.068 | 0.068 | 0.067 | 0.082 | 0.076 |
| 150 | - | 0.111 | 0.104 | 0.092 | 0.110 | 0.071 | 0.072 | 0.067 | 0.065 | 0.070 | 0.060 | 0.068 |
| 160 | - | 0.105 | 0.089 | 0.116 | 0.086 | 0.083 | 0.074 | 0.083 | 0.056 | 0.082 | 0.055 | 0.057 |
| 170 | - | 0.125 | 0.095 | 0.094 | 0.083 | 0.076 | 0.089 | 0.063 | 0.071 | 0.072 | 0.055 | 0.056 |
| 180 | - | 0.107 | 0.097 | 0.082 | 0.096 | 0.111 | 0.099 | 0.065 | 0.056 | 0.075 | 0.051 | 0.056 |
| 190 | - | 0.127 | 0.103 | 0.089 | 0.081 | 0.081 | 0.074 | 0.068 | 0.058 | 0.053 | 0.047 | 0.042 |
| 200 | - | 0.093 | 0.115 | 0.082 | 0.102 | 0.072 | 0.078 | 0.051 | 0.059 | 0.054 | 0.049 | 0.043 |

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05.

Table 3.6: Dividing Line D^* and D^{**} with N in Design (i)

| | | Sample size N | | | | | | | | | | |
|----------|-----|-----------------|-----|-----|-----|-----|-----|-----|-----|------|------|------|
| D | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| D^* | - | - | - | 170 | 140 | 90 | 80 | 70 | 80 | 70 | 40 | 40 |
| D^{**} | - | - | - | - | - | - | - | 200 | 200 | 190 | 160 | 160 |

Table 3.7: Dividing Line D^* and D^{**} with N in Design (ii)

| | | Sample size N | | | | | | | | | | |
|----------|-----|-----------------|-----|-----|-----|-----|-----|-----|-----|------|------|------|
| D | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| D^* | - | 200 | 190 | 170 | 160 | 90 | 80 | 80 | 70 | 70 | 60 | 50 |
| D^{**} | - | - | - | - | - | - | - | 200 | 200 | 190 | 160 | 160 |

Table 3.8: Dividing Line D^* and D^{**} with N in Design (iii)

| | | Sample size N | | | | | | | | | | |
|----------|-----|-----------------|-----|-----|-----|-----|-----|-----|-----|------|------|------|
| D | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| D^* | - | 200 | - | 170 | 160 | 90 | 70 | 90 | 80 | 70 | 70 | 60 |
| D^{**} | - | - | - | - | - | - | - | 200 | 200 | - | 170 | 170 |

Table 3.9: Dividing Line D^* and D^{**} with N in Design (iv)

| | | Sample size N | | | | | | | | | | |
|----------|-----|-----------------|-----|-----|-----|-----|-----|-----|-----|------|------|------|
| D | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| D^* | - | 200 | - | 170 | 160 | 90 | 80 | 80 | 70 | 80 | 80 | 80 |
| D^{**} | - | - | - | - | - | - | - | 200 | - | 200 | 180 | 180 |

Table 3.10: Dividing Line D^* and D^{**} with N in Design (v)

| | | Sample size N | | | | | | | | | | |
|----------|-----|-----------------|-----|-----|-----|-----|-----|-----|-----|------|------|------|
| D | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1500 | 2000 |
| D^* | - | - | - | 170 | - | 190 | 130 | 110 | 100 | 100 | 90 | 90 |
| D^{**} | - | - | - | - | - | - | - | - | - | - | 190 | 190 |

3.5 Conclusion

The main scenarios to be investigated steadily evolve into the “large dimension and large sample size” over the last 25 years. Not only do the techniques develop rapidly, but the high-dimensional theories also advance dramatically. Although many methods have been proposed for dimensional scenario; however, a more basic question about whether random vectors from a finite data sample is in high dimensionality or not, has not been considered in the literature. Similar to determining how large the sample size N is adequate to hold the CLT in a finite sample, one might ask how large the dimension D of the random vectors need to be for these high-dimensional theories to hold.

This chapter provides a general testing method to distinguish high dimensionality of random vectors from non-high. We also provide guidance to the global threshold for our test between high dimensionality and non-high for random vectors. The simulation study shows the size and power of our proposed test and illustrates that if we observed enough number of random vectors, only the underlying distribution of the random vectors in the finite sample alters the dimension that is needed for high dimensionality. If the underlying distribution is more complex, a larger dimension of random vectors is required for high dimensionality as the cost to achieve the high dimension properties. Further study for testing high dimensionality for the estimated vector in linear and non-linear regression is provided in the second part of our paper “Test for High Dimensionality of Random and Estimated Vectors”, co-authored with Tao Chen and Zetian Zhang.

References

- ALLEN-ZHU, Z., Y. LI, AND Z. SONG (2019): “A convergence theory for deep learning via over-parameterization,” in *International Conference on Machine Learning*, PMLR, 242–252.
- ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): “Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools,” *Journal of political economy*, 113, 151–184.
- ANDREWS, I., J. H. STOCK, AND L. SUN (2019): “Weak instruments in instrumental variables regression: Theory and practice,” *Annual Review of Economics*, 11, 727–753.
- ANGRIST, J. D., G. W. IMBENS, AND D. B. RUBIN (1996): “Identification of causal effects using instrumental variables,” *Journal of the American statistical Association*, 91, 444–455.
- ANTHONY, M., P. L. BARTLETT, P. L. BARTLETT, ET AL. (1999): *Neural network learning: Theoretical foundations*, vol. 9, cambridge university press Cambridge.
- BAI, Z. AND H. SARANADASA (1996): “Effect of high dimension: by an example of a two sample problem,” *Statistica Sinica*, 311–329.
- BARTAN, B. AND M. PILANCI (2021): “Neural spectrahedra and semidefinite lifts: Global convex optimization of polynomial activation neural networks in fully polynomial-time,” *arXiv preprint arXiv:2101.02429*.
- BELLMAN, R. AND R. KALABA (1957): “Dynamic programming and statistical communication theory,” *Proceedings of the National Academy of Sciences*, 43, 749–751.

- BENNETT, A., N. KALLUS, AND T. SCHNABEL (2019): “Deep generalized method of moments for instrumental variable analysis,” *Advances in neural information processing systems*, 32.
- BLACKWELL, M. (2014): “A selection bias approach to sensitivity analysis for causal effects,” *Political Analysis*, 22, 169–182.
- BONVINI, M. AND E. H. KENNEDY (2022): “Sensitivity analysis via the proportion of unmeasured confounding,” *Journal of the American Statistical Association*, 117, 1540–1550.
- BOTTOU, L., F. E. CURTIS, AND J. NOCEDAL (2018): “Optimization methods for large-scale machine learning,” *Siam Review*, 60, 223–311.
- BRUMBACK, B. A., M. A. HERNÁN, S. J. HANEUSE, AND J. M. ROBINS (2004): “Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures,” *Statistics in medicine*, 23, 749–767.
- BUHRMESTER, V., D. MÜNCH, AND M. ARENS (2021): “Analysis of explainers of black box deep neural networks for computer vision: A survey,” *Machine Learning and Knowledge Extraction*, 3, 966–989.
- CALHOUN, G. (2011): “Hypothesis testing in linear regression when k/n is large,” *Journal of econometrics*, 165, 163–174.
- CARD, D. (1999): “The causal effect of education on earnings,” *Handbook of labor economics*, 3, 1801–1863.
- CARNEGIE, N. B., M. HARADA, AND J. L. HILL (2016): “Assessing sensitivity to unmeasured confounding using a simulated potential confounder,” *Journal of Research on Educational Effectiveness*, 9, 395–420.
- CHAO, J. C. AND N. R. SWANSON (2005): “Consistent estimation with a large number of weak instruments,” *Econometrica*, 73, 1673–1692.
- CHEN, J., D. L. CHEN, AND G. LEWIS (2020): “Mostly harmless machine learning: learning optimal instruments in linear IV models,” *arXiv preprint arXiv:2011.06158*.

- CHEN, J., X. CHEN, AND E. TAMER (2023): “Efficient estimation of average derivatives in NPIV models: Simulation comparisons of neural network estimators,” *Journal of Econometrics*, 235, 1848–1875.
- CHEN, X. (2007): “Large sample sieve estimation of semi-nonparametric models,” *Handbook of econometrics*, 6, 5549–5632.
- CHEN, X. AND Z. LIAO (2015): “Sieve semiparametric two-step GMM under weak dependence,” *Journal of Econometrics*, 189, 163–186.
- CHEN, X. AND X. SHEN (1998): “Sieve extremum estimates for weakly dependent data,” *Econometrica*, 289–314.
- CHERNOZHUKOV, V., D. CHETVERIKOV, K. KATO, ET AL. (2017): “Central limit theorems and bootstrap in high dimensions,” *Annals of Probability*, 45, 2309–2352.
- CHERNOZHUKOV, V., C. CINELLI, W. NEWEY, A. SHARMA, AND V. SYRGKANIS (2022): “Long story short: Omitted variable bias in causal machine learning,” Tech. rep., National Bureau of Economic Research.
- CINELLI, C. AND C. HAZLETT (2020): “Making sense of sensitivity: Extending omitted variable bias,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 39–67.
- CONNEAU, A., H. SCHWENK, L. BARRAULT, AND Y. LECUN (2016): “Very deep convolutional networks for natural language processing,” *arXiv preprint arXiv:1606.01781*, 2.
- CORNFIELD, J., W. HAENSZEL, E. C. HAMMOND, A. M. LILIENFELD, M. B. SHIMKIN, AND E. L. WYNDER (1959): “Smoking and lung cancer: recent evidence and a discussion of some questions,” *Journal of the National Cancer institute*, 22, 173–203.
- DANGETI, P. (2017): *Statistics for machine learning*, Packt Publishing Ltd.
- DAVIDSON, R. AND J. G. MACKINNON (2006): “The power of bootstrap and asymptotic tests,” *Journal of Econometrics*, 133, 421–441.

- DE BOOR, C. AND C. DE BOOR (1978): *A practical guide to splines*, vol. 27, springer-verlag New York.
- DEMPSTER, A. P. (1958): “A high dimensional two sample significance test,” *The Annals of Mathematical Statistics*, 995–1010.
- DOKSUM, K. AND A. SAMAROV (1995): “Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression,” *The Annals of Statistics*, 1443–1473.
- DONG, W., P. WANG, W. YIN, G. SHI, F. WU, AND X. LU (2018): “Denoising prior driven deep neural network for image restoration,” *IEEE transactions on pattern analysis and machine intelligence*, 41, 2305–2318.
- DONOHO, D. L. ET AL. (2000): “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS math challenges lecture*, 1, 32.
- DORIE, V., M. HARADA, N. B. CARNEGIE, AND J. HILL (2016): “A flexible, interpretable framework for assessing sensitivity to unmeasured confounding,” *Statistics in medicine*, 35, 3453–3470.
- EMMONS, J., S. FOULADI, G. ANANTHANARAYANAN, S. VENKATARAMAN, S. SAVARESE, AND K. WINSTEIN (2019): “Cracking open the dnn black-box: Video analytics with dnns across the camera-cloud boundary,” in *Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges*, 27–32.
- FABOZZI, F. J., H. FALLAHGOUL, V. FRANSTIANTO, AND G. LOEPER (2019): “Towards explaining deep learning: Asymptotic properties of relu ffn sieve estimators,” *Available at SSRN 3499324*.
- FARRELL, M. H., T. LIANG, AND S. MISRA (2021): “Deep neural networks for estimation and inference,” *Econometrica*, 89, 181–213.
- FISHER, R. A. AND L. H. C. TIPPETT (1928): “Limiting forms of the frequency distribution of the largest or smallest member of a sample,” in *Mathematical proceedings of the Cambridge philosophical society*, Cambridge University Press, vol. 24, 180–190.

- FOHR, D., O. MELLA, AND I. ILLINA (2017): “New paradigm in speech recognition: deep neural networks,” in *IEEE international conference on information systems and economic intelligence*.
- FORCHINI, G. AND B. JIANG (2019): “The unconditional distributions of the OLS, TSLS and LIML estimators in a simple structural equations model,” *Econometric Reviews*, 38, 208–247.
- FRANK, K. A. (2000): “Impact of a confounding variable on a regression coefficient,” *Sociological Methods & Research*, 29, 147–194.
- FRANK, K. A., S. J. MAROULIS, M. Q. DUONG, AND B. M. KELCEY (2013): “What would it take to change an inference? Using Rubin’s causal model to interpret the robustness of causal inferences,” *Educational Evaluation and Policy Analysis*, 35, 437–460.
- GALLANT (1988): “There exists a neural network that does not make avoidable mistakes,” in *IEEE 1988 International Conference on Neural Networks*, IEEE, 657–664.
- GOODFELLOW, I., Y. BENGIO, AND A. COURVILLE (2016): *Deep learning*, MIT press.
- GRENANDER, U. (1981): “Abstract inference,” (*No Title*).
- GROMOV, M. AND V. D. MILMAN (1983): “A topological application of the isoperimetric inequality,” *American Journal of Mathematics*, 105, 843–854.
- HAINMUELLER, J. AND C. HAZLETT (2014): “Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach,” *Political Analysis*, 22, 143–168.
- HALL, A. R., G. D. RUDEBUSCH, AND D. W. WILCOX (1996): “Judging instrument relevance in instrumental variables estimation,” *International Economic Review*, 283–298.
- HANSEN, C., J. HAUSMAN, AND W. NEWEY (2008): “Estimation with many instrumental variables,” *Journal of Business & Economic Statistics*, 26, 398–422.

- HARTFORD, J., G. LEWIS, K. LEYTON-BROWN, AND M. TADDY (2017): “Deep IV: A flexible approach for counterfactual prediction,” in *International Conference on Machine Learning*, PMLR, 1414–1423.
- HE, X. AND Q.-M. SHAO (2000): “On parameters of increasing dimensions,” *Journal of multivariate analysis*, 73, 120–135.
- HECKMAN, J. (1997): “Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations,” *Journal of human resources*, 441–462.
- HECKMAN, J. AND S. NAVARRO-LOZANO (2004): “Using matching, instrumental variables, and control functions to estimate economic choice models,” *Review of Economics and statistics*, 86, 30–57.
- HECKMAN, J. J. AND E. J. VYTLACIL (2001): “Instrumental variables, selection models, and tight bounds on the average treatment effect,” in *Econometric Evaluation of Labour Market Policies*, Springer, 1–15.
- HOREL, E. AND K. GIESECKE (2020): “Significance tests for neural networks,” *Journal of Machine Learning Research*, 21, 1–29.
- HOROWITZ, J. L. (2019): “Bootstrap methods in econometrics,” *Annual Review of Economics*, 11, 193–224.
- HOSMAN, C. A., B. B. HANSEN, AND P. W. HOLLAND (2010): “The sensitivity of linear regression coefficients’ confidence limits to the omission of a confounder,” .
- HUBER, P. J. (1973): “Robust regression: asymptotics, conjectures and Monte Carlo,” *The annals of statistics*, 799–821.
- HUSMEIER, D. (2012): *Neural networks for conditional probability estimation: Forecasting beyond point predictions*, Springer Science & Business Media.
- IMAI, K., L. KEELE, AND T. YAMAMOTO (2010): “Identification, inference and sensitivity analysis for causal mediation effects,” .
- IMBENS, G. W. (2003): “Sensitivity to exogeneity assumptions in program evaluation,” *American Economic Review*, 93, 126–132.

- IMBENS, G. W. AND D. B. RUBIN (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.
- JESSON, A., S. MINDERMANN, Y. GAL, AND U. SHALIT (2021): “Quantifying ignorance in individual-level causal-effect estimates under hidden confounding,” in *International Conference on Machine Learning*, PMLR, 4829–4838.
- KALLUS, N. (2023): “Treatment effect risk: Bounds and inference,” *Management Science*, 69, 4579–4590.
- KALLUS, N., X. MAO, AND A. ZHOU (2019): “Interval estimation of individual-level causal effects under unobserved confounding,” in *The 22nd international conference on artificial intelligence and statistics*, PMLR, 2281–2290.
- KALLUS, N. AND A. ZHOU (2018): “Confounding-robust policy improvement,” *Advances in neural information processing systems*, 31.
- LEE, D. S., J. MCCRARY, M. J. MOREIRA, AND J. R. PORTER (2021): “Valid t-ratio Inference for IV,” Tech. rep., National Bureau of Economic Research.
- LEE, L.-F. (2003): “Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances,” *Econometric Reviews*, 22, 307–335.
- LEHMANN, E. L. (1999): *Elements of large-sample theory*, Springer.
- LEWIS, G. AND V. SYRGKANIS (2018): “Adversarial generalized method of moments,” *arXiv preprint arXiv:1803.07164*.
- LINDSAY, B. G., J. KETTENRING, AND D. O. SIEGMUND (2004): “A report on the future of statistics,” .
- LIU, R., Z. SHANG, AND G. CHENG (2020): “On deep instrumental variables estimate,” *arXiv preprint arXiv:2004.14954*.
- MATHEW, A., P. AMUDHA, AND S. SIVAKUMARI (2021): “Deep learning techniques: an overview,” *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, 599–608.

- MIDDLETON, J. A., M. A. SCOTT, R. DIAKOW, AND J. L. HILL (2016): “Bias amplification and bias unmasking,” *Political Analysis*, 24, 307–323.
- OSTER, E. (2019): “Unobservable selection and coefficient stability: Theory and evidence,” *Journal of Business & Economic Statistics*, 37, 187–204.
- PEARL, J. (1995): “Causal inference from indirect experiments,” *Artificial intelligence in medicine*, 7, 561–582.
- PEARSON, K. (1905): *On the general theory of skew correlation and non-linear regression*, 14, Dulau and Company.
- PEROLAT, J., B. DE VYLDER, D. HENNES, E. TARASSOV, F. STRUB, V. DE BOER, P. MULLER, J. T. CONNOR, N. BURCH, T. ANTHONY, ET AL. (2022): “Mastering the game of stratego with model-free multiagent reinforcement learning,” *Science*, 378, 990–996.
- PORTNOY, S. (1984): “Asymptotic behavior of M -estimators of p regression parameters when p^2/n is large. I. Consistency,” *The Annals of Statistics*, 1298–1309.
- PORTNOY, STEPHEN (1985): “Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation,” *The Annals of Statistics*, 13, 1403–1417.
- PORTNOY, STEPHEN (1988): “Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity,” *The annals of statistics*, 356–366.
- RAWAT, W. AND Z. WANG (2017): “Deep convolutional neural networks for image classification: A comprehensive review,” *Neural computation*, 29, 2352–2449.
- RIEGG, S. K. (2008): “Causal inference and omitted variable bias in financial aid research: Assessing solutions,” *The Review of Higher Education*, 31, 329–354.
- ROBINS, J. M. (1999): “Association, causation, and marginal structural models,” *Synthese*, 121, 151–179.

- ROSENBAUM, P. R. AND P. R. ROSENBAUM (2002): *Overt bias in observational studies*, Springer.
- ROSENBAUM, P. R. AND D. B. RUBIN (1983): “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 45, 212–218.
- ROSENBAUM, PAUL R AND RUBIN, DONALD B (1983): “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70, 41–55.
- ROWELL, J. AND D. WALTERS (1976): “Analysing data with repeated observations on each experimental unit,” *The Journal of Agricultural Science*, 87, 423–432.
- SANDERSON, E. AND F. WINDMEIJER (2016): “A weak instrument F-test in linear IV models with multiple endogenous variables,” *Journal of econometrics*, 190, 212–221.
- SCHARFSTEIN, D. O., R. NABI, E. H. KENNEDY, M.-Y. HUANG, M. BONVINI, AND M. SMID (2021): “Semiparametric sensitivity analysis: Unmeasured confounding in observational studies,” *arXiv preprint arXiv:2104.08300*.
- SCHUMAKER, L. (2007): *Spline functions: basic theory*, Cambridge university press.
- SHEN, X., C. JIANG, L. SAKHANENKO, AND Q. LU (2023): “Asymptotic properties of neural network sieve estimators,” *Journal of nonparametric statistics*, 35, 839–868.
- SHEN, XIAOXI AND JIANG, CHANG AND SAKHANENKO, LYUDMILA AND LU, QING (2019): “Asymptotic properties of neural network sieve estimators,” *arXiv preprint arXiv:1906.00875*.
- SHEU, Y.-H. (2020): “Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research,” *Frontiers in Psychiatry*, 11.
- STAIGER, D. AND J. H. STOCK (1997): “Instrumental variables regression with weak instruments,” *Econometrica*, 65, 557–586.
- STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): “A survey of weak instruments and weak identification in generalized method of moments,” *Journal of Business & Economic Statistics*, 20, 518–529.

- STOCK, J. H. AND M. YOGO (2002): “Testing for weak instruments in linear IV regression,” .
- SUR, P., Y. CHEN, AND E. J. CANDÈS (2019): “The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square,” *Probability theory and related fields*, 175, 487–558.
- VANDERWEELE, T. J. AND O. A. ARAH (2011): “Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders,” *Epidemiology*, 22, 42–52.
- VANDERWEELE, T. J. AND P. DING (2017): “Sensitivity analysis in observational research: introducing the E-value,” *Annals of internal medicine*, 167, 268–274.
- VERSHYNIN, R. (2020): “High-dimensional probability,” *University of California, Irvine*, 10, 11.
- WILMS, R., E. MÄTHNER, L. WINNEN, AND R. LANWEHR (2021): “Omitted variable bias: A threat to estimating causal relationships,” *Methods in Psychology*, 5, 100075.
- WÜTHRICH, K. AND Y. ZHU (2023): “Omitted variable bias of Lasso-based inference methods: A finite sample analysis,” *Review of Economics and Statistics*, 105, 982–997.
- ZENG, J., T. T.-K. LAU, S. LIN, AND Y. YAO (2019): “Global convergence of block coordinate descent in deep learning,” in *International Conference on Machine Learning*, PMLR, 7313–7323.
- ZHAO, W. X., K. ZHOU, J. LI, T. TANG, X. WANG, Y. HOU, Y. MIN, B. ZHANG, J. ZHANG, Z. DONG, ET AL. (2023): “A survey of large language models,” *arXiv preprint arXiv:2303.18223*.

APPENDICES

A Appendices of Chapter 1

This appendix contains proofs to lemmas and theorems. The proofs apply the properties of the Frobenius norm (hereafter, the F-prop.), as such

- (a). (subadditivity) $\|A + B\|_F \leq \|A\|_F + \|B\|_F$, for all finite dimensional matrices A and B of size d_1 -by- d_2 ;
- (b). (submultiplicativity) $\|AB\|_F \leq \|A\|_F \|B\|_F$, for all finite dimensional matrices A of size d_1 -by- d_2 and B of size d_2 -by- d_3 ;
- (c). (trace inequality) $|\text{tr}(A)| \leq \sqrt{d}\|A\|_F$ for a d -by- d matrix A .

Lemma A.1. *Suppose Assumptions 1.2.1 to 1.2.3 hold. Applying Lemma 1.2.2, we have*

- (a). $\|n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W})X_i^\top - \Sigma_{gX}\|_F = o_p(1)$;
- (b). $\|\{n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W})X_i^\top\}^{-1} - \Sigma_{gX}^{-1}\|_F = o_p(1)$.

Lemma A.2. *Let f be a function that maps a squared matrix to a real value; then for full rank squared matrices A , B and C of a dimension d -by- d , there is*

$$f(A) = f(B) + \text{tr} \left[\left\{ \frac{\partial f(C)}{\partial C} \right\}^\top (A - B) \right],$$

where $\min\{a_{ij}, b_{ij}\} < c_{ij} < \max\{a_{ij}, b_{ij}\}$ for all elements a_{ij} , b_{ij} and c_{ij} of the matrices A , B and C , respectively, for $i, j = 1, \dots, d$.

A.1 Proof of Theorems

Now we prove the theorems in Sections 1.2 and 1.3.

Proof of Theorem 1.2.1

We have

$$\|\hat{\beta}_{\text{DNN}} - \beta_0\|_F^2 = \left\| \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i \right\} \right\|_F^2 \quad ((1.2.1) \ \& \ (1.2.5))$$

$$\stackrel{(F\text{-prop.})}{\leq} \left\| \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top \right\}^{-1} - \Sigma_{gX}^{-1} \right\|_F^2 \left\| \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i \right\|_F^2 + \quad (\text{A.1})$$

$$\left\| \Sigma_{gX}^{-1} \right\|_F^2 \left\| \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i \right\|_F^2. \quad (\text{A.2})$$

Provided $\|n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i\|_F^2 = o_p(1)$, there is $\|\hat{\beta}_{\text{DNN}} - \beta_0\|_F^2 \xrightarrow{p} 0$, because (A.1) = $o_p(1)$ by Lemma A.1(b) and (A.2) = $o_p(1)$ by Assumption 1.2.3(b). It then remains to show that $\|n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i\|_F^2 = o_p(1)$.

Note that the composition of polynomials is still a polynomial; hence, the network built with polynomial activation functions can be expressed as $s(Z_i) = \sum_{m=1}^{d_P} [C_W]_{\cdot, m} P^{(m)}(Z_i) = C_W P(Z_i)$. In the above expression, $P^{(m)}(Z_i)$ denotes the m th entry of vector $P(Z_i)$, which is an d_P -vector consisting of the power functions of each input IV as well as their interaction terms. For example, if the polynomials are of degree 2 and there are two IVs as such $Z_i = [Z_{i1}, Z_{i2}]^\top$, then $P(Z_i) = [1, Z_{i1}, Z_{i2}, Z_{i1}^2, Z_{i2}^2, Z_{i1}Z_{i2}, Z_{i1}^2Z_{i2}, Z_{i1}Z_{i2}^2, Z_{i1}^2Z_{i2}^2]^\top$ with $d_P = 9$. $[C_W]_{\cdot, m}$ denotes the m th column of a d_X -by- d_P coefficient matrix C_W , which contains various transformations of all the weights W . Hence, with the trained weights \hat{W} , we have $g_n^s(Z_i; \hat{W}) = \sum_{m=1}^{d_P} [C_{\hat{W}}]_{\cdot, m} P^{(m)}(Z_i) = C_{\hat{W}} P(Z_i)$, and thus,

$$\frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i = \sum_{m=1}^{d_P} [C_{\hat{W}}]_{\cdot, m} \frac{1}{n} \sum_{i=1}^n \{P^{(m)}(Z_i) U_i\} = C_{\hat{W}} \frac{1}{n} \sum_{i=1}^n \{P(Z_i) U_i\}.$$

Then, given any degree \bar{K} of the Bernstein polynomial approximation and its approximating network s , as $n \rightarrow \infty$, there is $\|n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i\|_F^2 \leq \|C_{\hat{W}}\|_F^2 \|n^{-1} \sum_{i=1}^n \{P(Z_i) U_i\}\|_F^2 =$

$o_p(1)$, because $\|C_{\hat{W}}\|_F^2 = \mathcal{O}_p(1)$, and $\|n^{-1} \sum_{i=1}^n \{P(Z_i)U_i\}\|_F^2 = o_p(1)$ by Assumption 1.2.2. \square

Proof of Theorem 1.2.2

Under the structural model (1.2.1), we have

$$\sqrt{n}(\hat{\beta}_{\text{DNN}} - \beta_0) = \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i \right\}. \quad (\text{A.3})$$

Note that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) U_i = \sum_{m=1}^{d_P} [C_{\hat{W}}]_{\cdot, m} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{P^{(m)}(Z_i) U_i\} = C_{\hat{W}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{P(Z_i) U_i\}.$$

For any given degree \bar{K} of the Bernstein polynomial approximation and its approximating network s , provided the training converges as the rounds of iteration $t \rightarrow \infty$ for each n such that $\|C_{\hat{W}^{(t)}} P(Z_i) - C_W P(Z_i)\| \xrightarrow{p} 0$ for all Z_i (see, e.g., Allen-Zhu et al., 2019; Zeng et al., 2019), it follows that $\|C_{\hat{W}^{(t)}} - C_W\| \xrightarrow{p} 0$ as $t \rightarrow \infty$ for each n as $n \rightarrow \infty$, where the terms with the superscript (t) indicates the quantities in the t th learning iteration. Meanwhile, it is implied by Assumption 1.2.2 that $\mathbb{E}[P(Z_i)U_i] = 0$ and by Assumption 1.2.5 that $\mathbb{E}[(P^{(m)}(Z_i)U_i)^2] < \infty$. Let G_i denote a centered Gaussian random vector of the same dimension as $P(Z_i)U_i$ such that $G_i \sim \mathcal{N}(0, \mathbb{E}[P(Z_i)U_i^2 P^\top(Z_i)])$ for all i . Then applying Theorem 2.1 from Chernozhukov et al. (2017) and our Assumption 1.2.4 yields $err_n(\mathcal{A}) := \sup_{A \in \mathcal{A}} |\mathbb{P}(S_n \in A) - \mathbb{P}(S_n^G \in A)| \rightarrow 0$ with a proper speed of $d_P \rightarrow \infty$ as $n \rightarrow \infty$, where \mathcal{A} is a class of Borel sets in \mathbb{R}^{d_P} , $S_n := n^{-1/2} \sum_{i=1}^n \{P(Z_i)U_i\}$ and $S_n^G := n^{-1/2} \sum_{i=1}^n G_i$. Hence, by Slutsky's theorem and the continuous mapping theorem, we have

$$err_n(\mathcal{B}) := \sup_{B \in \mathcal{B}} |\mathbb{P}(R_n \in B) - \mathbb{P}(R_n^G \in B)| \rightarrow 0$$

as $t \rightarrow \infty$ for $n \rightarrow \infty$, where \mathcal{B} is a class of Borel sets in \mathbb{R}^{d_X} , $R_n^G := \Sigma_{BX}^{-1} (C_W S_n^G)$ whence $R_n^G \sim \mathcal{N}(0, \sigma_{UU} \Sigma_{BX}^{-1} \Sigma_B \Sigma_{BX}^{-\top})$, and $R_n := \{n^{-1} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top\}^{-1} C_{\hat{W}} S_n$. The variance

$\text{Var}(R_n)$ can then be estimated as

$$\text{Var}(\hat{R}_n) := \left(\frac{1}{n} \sum_{j=1}^n \hat{X}_j X_j^\top \right)^{-1} \left(\frac{1}{n} \sum_{j=1}^n \hat{X}_j \hat{U}_j^2 \hat{X}_j^\top \right) \left(\frac{1}{n} \sum_{j=1}^n X_j \hat{X}_j^\top \right)^{-1},$$

where $\hat{X}_i := g_n^s(Z_i; \hat{W})$ and $\hat{U}_i := Y_i - X_i^\top \hat{\beta}_{\text{DNN}}$. This estimated variance is consistent due to $(n^{-1} \sum_{j=1}^n \hat{X}_j X_j^\top)^{-1} \xrightarrow{p} \Sigma_{BX}^{-1}$ by Lemma A.1(b) and $n^{-1} \sum_{j=1}^n \hat{X}_j \hat{U}_j^2 \hat{X}_j^\top \xrightarrow{p} \sigma_{UU} \Sigma_B$ by $\|C_{\hat{W}(t)} P(Z_i) - C_W P(Z_i)\| \xrightarrow{p} 0$ for all Z_i as $t \rightarrow \infty$. \square

Proof of Theorem 1.2.3

With some $\gamma_n^s(Z_i; \hat{W})$ satisfying Lemma 1.2.2 for γ , Assumption 1.2.6(c) implies that

$$\frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \beta)\} \xrightarrow{p} \mathbb{E}[\gamma(Z_i) \{Y_i - f(X_i, \beta)\}].$$

Since $\mathbb{E}[\gamma(Z_i) \{Y_i - f(X_i, \beta)\}]$ has a unique zero point at $\beta = \beta_0$ by Assumption 1.2.6(b), the solution $\tilde{\beta}_{\text{DNN}}$ by solving (1.2.6) is such that $\tilde{\beta}_{\text{DNN}} \xrightarrow{p} \beta_0$ by Assumption 1.2.6(a). \square

Proof of Theorem 1.2.4

By the MVT, there is

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \tilde{\beta}_{\text{DNN}})\} \\ &= \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \beta_0)\} + \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \nabla_{\beta} f(X_i, \bar{\beta}) (\tilde{\beta}_{\text{DNN}} - \beta_0), \end{aligned}$$

with some $\bar{\beta}$, such that $\bar{\beta}^{(d)} \in [\tilde{\beta}_{\text{DNN}}^{(d)}, \beta_0^{(d)}]$ for all $d = 1, \dots, d_X$. Then, under the first order condition and Assumptions 1.2.7(a) and 1.2.7(b), we have

$$\sqrt{n}(\tilde{\beta}_{\text{DNN}} - \beta_0) = - \left\{ \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \nabla_{\beta} f(X_i, \bar{\beta}) \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \beta_0)\}$$

Provided the consistency of $\tilde{\beta}_{\text{DNN}}$, it is implied that $\bar{\beta} \xrightarrow{p} \beta_0$. Hence, with some $\gamma_n^s(Z_i; \hat{W})$ satisfying Lemma 1.2.2 for γ ,

$$\left\{ \frac{1}{n} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \nabla_{\beta} f(X_i, \bar{\beta}) \right\}^{-1} \xrightarrow{p} \mathbb{E}[B_K(Z_i, \gamma) \{\nabla_{\beta} f(X_i, \beta_0)\}^{\top}]^{-1},$$

under Assumption 1.2.7(c) and Slutsky's theorem. Meanwhile, with such $\gamma_n^s(Z_i; \hat{W})$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \gamma_n^s(Z_i; \hat{W}) \{Y_i - f(X_i, \beta_0)\} \xrightarrow{d} N(0, \Sigma_{\gamma U})$$

under the structural model (1.2.1*) and Assumption 1.2.7(d). Therefore, it follows that

$$\begin{aligned} \sqrt{n} (\tilde{\beta}_{\text{DNN}} - \beta_0) &\xrightarrow{d} \\ &\mathcal{N}\left(0, \mathbb{E}[B_K(Z_i, \gamma) \{\nabla_{\beta} f(X_i, \beta_0)\}^{\top}]^{-1} \Sigma_{\gamma U} \mathbb{E}[\nabla_{\beta} f(X_i, \beta_0) \{B_K(Z_i, \gamma)\}^{\top}]^{-1}\right). \end{aligned}$$

□

Proof of Theorem 1.3.1

First, we show that under the null hypothesis $H_0 : g = 0$, $p_L GF \xrightarrow{d} \chi_{p_L}^2$. Since the network is a composition as shown in (1.2.4), the first-order condition of the loss function in the first stage implies

$$\mathbf{0} = \nabla_{W_L} \text{Loss}_n^s(\hat{W}) = \frac{1}{n} \sum_{i=1}^n \{X_i - g_n^s(Z_i; \hat{W})\} \hat{h}_L^{\top}(Z_i) = \frac{1}{n} \sum_{i=1}^n \{X_i - \hat{W}_L \hat{h}_L(Z_i)\} \hat{h}_L^{\top}(Z_i),$$

where $\hat{h}_{i,1} := \text{poly}(\hat{W}_0 Z_i)$, and $\hat{h}_{i,l+1} := \text{poly}(\hat{W}_l \hat{h}_{i,l})$ for $i = 1, \dots, n$ and $l = 1, \dots, L-1$. Solving the first order condition using the stochastic gradient descent algorithm, it follows that given any proper sample size n ,

$$\frac{1}{n} \sum_{i=1}^n \{X_i - \hat{W}_L^{(t)} \hat{h}_L^{(t)}(Z_i)\} \hat{h}_L^{(t)\top}(Z_i) \xrightarrow{p} \mathbf{0}, \text{ as } t \rightarrow \infty,$$

where the terms with the superscript (t) indicates the quantities in the t th learning iteration (Bottou et al., 2018), and thus, as $t \rightarrow \infty$,

$$\left\| \frac{1}{n} \sum_{i=1}^n \left\{ X_i \hat{h}_L^{(t)\top}(Z_i) \right\} \left[\frac{1}{n} \sum_{i=1}^n \left\{ \hat{h}_L^{(t)}(Z_i) \hat{h}_L^{(t)\top}(Z_i) \right\} \right]^{-1} - \hat{W}_L^{(t)} \right\|_{\max} = o_p(1).$$

It is then implied that for any $\varepsilon > 0$, there exists some $T(\varepsilon) > 0$ such that for all $t > T(\varepsilon)$,

$$\begin{aligned} g_n^s(Z_j; \hat{W}) &= \hat{W}_L^{(t)} \hat{h}_L^{(t)}(Z_j) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ X_i \hat{h}_L^{(t)\top}(Z_i) \right\} \left[\frac{1}{n} \sum_{i=1}^n \left\{ \hat{h}_L^{(t)}(Z_i) \hat{h}_L^{(t)\top}(Z_i) \right\} \right]^{-1} \hat{h}_L^{(t)}(Z_j) + \varepsilon \iota_W \hat{h}_L^{(t)}(Z_j), \end{aligned}$$

where ι_W denotes a matrix of value 1 of the same dimension as $\hat{W}_L^{(t)}$. Hence,

$$\begin{aligned} &\sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ X_i \hat{h}_L^{(t)\top}(Z_i) \right\} \left[\frac{1}{n} \sum_{i=1}^n \left\{ \hat{h}_L^{(t)}(Z_i) \hat{h}_L^{(t)\top}(Z_i) \right\} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \hat{h}_L^{(t)}(Z_i) X_i^\top \right\} + \mathcal{O}_p(n\varepsilon), \end{aligned}$$

since $\varepsilon \iota_W \sum_{i=1}^n \left\{ \hat{h}_L^{(t)}(Z_i) X_i^\top \right\} = \mathcal{O}_p(n\varepsilon)$. With polynomial activation functions, the term $\hat{h}_L^{(t)}(Z_i)$ can be expressed in a linear form as such $\hat{h}_L^{(t)}(Z_i) := C_{\hat{W},L} P_L(Z_i)$, where $P_L(Z_i)$ is a vector consisting of the power functions of each input IV as well as their interaction terms, similar to $P(Z_i)$, and $C_{\hat{W},L}$ is defined in the same way as $C_{\hat{W}}$ but only with the trained weights up to the L th hidden layer. Hence, one can choose a $T(\varepsilon)$ that is large enough such that $\mathcal{O}_p(n\varepsilon) = o_p(1)$, and applying the Slutsky's theorem yields

$$\begin{aligned} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ X_i P_L^\top(Z_i) \right\} C_{\hat{W},L}^\top \left[C_{\hat{W},L} \frac{1}{n} \sum_{i=1}^n \left\{ P_L(Z_i) P_L^\top(Z_i) \right\} C_{\hat{W},L}^\top \right]^{-1} \\ &\quad \frac{1}{\sqrt{n}} \sum_{i=1}^n C_{\hat{W},L} \left\{ P_L(Z_i) X_i^\top \right\} + o_p(1). \end{aligned}$$

Under the H_0 it follows that $X_i = V_i$, and thus,

$$\begin{aligned} & \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{V_i P_L^\top(Z_i)\} C_{\hat{W},L}^\top \left[C_{\hat{W},L} \frac{1}{n} \sum_{i=1}^n \{P_L(Z_i) P_L^\top(Z_i)\} C_{\hat{W},L}^\top \right]^{-1} \\ & \quad C_{\hat{W},L} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{P_L(Z_i) V_i^\top\} + o_p(1). \end{aligned} \tag{A.4}$$

By Assumptions 1.2.2 and 1.2.3, it is implied that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \{P_L(Z_i) V_i^\top\} \xrightarrow{d} N(0, \Sigma_{ZV}),$$

where $\Sigma_{ZV} := \mathbb{E}[P_L(Z_i) \sigma_{VV} P_L^\top(Z_i)] = \sigma_{VV} \mathbb{E}[P_L(Z_i) P_L^\top(Z_i)]$, whence by the Slutsky's theorem, we have

$$C_{\hat{W},L} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{P_L(Z_i) V_i^\top\} \xrightarrow{d} N(0, C_{W,L} \Sigma_{ZV} C_{W,L}^\top).$$

Under Assumptions 1.2.1 and 1.2.3(b), applying the continuous mapping theorem and Slutsky's theorem again, it follows that

$$\left[C_{\hat{W},L} \frac{1}{n} \sum_{i=1}^n \{P_L(Z_i) P_L^\top(Z_i)\} C_{\hat{W},L}^\top \right]^{-1} \xrightarrow{p} (C_{W,L} \sigma_{VV}^{-1} \Sigma_{ZV} C_{W,L}^\top)^{-1}.$$

Hence,

$$\frac{\sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top}{\sigma_{VV}} \xrightarrow{d} \chi_{pL}^2$$

Meanwhile, since $\hat{V}_i = X_i - g_n^s(Z_i; \hat{W}) = V_i + g(Z_i) - g_n^s(Z_i; \hat{W})$, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \hat{V}_i^2 &= \frac{1}{n} \sum_{j=1}^n \{V_j + g(Z_j) - g_n^s(Z_j; \hat{W})\}^2 + \mathcal{O}_p(\varepsilon) \\ &= \frac{1}{n} \sum_{j=1}^n V_j^2 + \frac{2}{n} \sum_{j=1}^n V_j \{g(Z_j) - g_n^s(Z_j; \hat{W})\} + \frac{1}{n} \sum_{j=1}^n \{g(Z_j) - g_n^s(Z_j; \hat{W})\}^2 + \mathcal{O}_p(\varepsilon) \\ &= \frac{1}{n} \sum_{j=1}^n V_j^2 + o_p(1). \end{aligned}$$

By Assumptions 1.2.1 and 1.2.2, $n^{-1} \sum_{j=1}^n V_j^2 \xrightarrow{p} \sigma_{VV}$. Therefore,

$$p_L GF = \left\{ \frac{\sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top}{\sigma_{VV}} \right\} \left(\frac{\sigma_{VV}}{n^{-1} \sum_{i=1}^n \hat{V}_i^2} \right) \xrightarrow{d} \chi_{p_L}^2.$$

Under the local alternatives of $H_a : g = o_p(n^{-1/2})$, $X_i = o_p(n^{-1/2}) + V_i$, and Equation (A.4) still holds. Thus, we still have $p_L GF \xrightarrow{d} \chi_{p_L}^2$. While under the local alternatives of $H_a : g \sim \mathcal{O}_p(n^{-1/2})$, $X_i = \mathcal{O}_p(n^{-1/2}) + V_i$, and thus $n^{-1/2} \sum_{i=1}^n \{P_L(Z_i) X_i^\top\} = n^{-1/2} \sum_{i=1}^n \{P_L(Z_i) V_i^\top\} + n^{-1/2} \sum_{i=1}^n \{P_L(Z_i) g(Z_i)\} \xrightarrow{d} N(\mu_{nc}, \Sigma_{ZV})$, where the noncentral expectation μ_{nc} is as such $\mu_{nc} := \mathbb{E}[P_L(Z_i) g(Z_i)]$. Hence, there is

$$C_{\hat{W},L} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{P_L(Z_i) X_i^\top\} \xrightarrow{d} N(C_{W,L} \mu_{nc}, C_{W,L} \Sigma_{ZV} C_{W,L}^\top),$$

and thus,

$$\frac{\sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top}{\sigma_{VV}} \xrightarrow{d} \chi_{p_L}^2(\alpha),$$

where $\alpha = p_L \mu_{nc}^\top C_{W,L}^\top C_{W,L} \mu_{nc}$. Yet under the alternatives of $H_a : g \notin \mathcal{O}_p(n^{-1/2})$, we have $\sum_{i=1}^n \{g_n^s(Z_i; \hat{W}) X_i^\top\}$ diverging with n , while $n^{-1} \sum_{i=1}^n \hat{V}_i^2 = \mathcal{O}_p(1)$. Hence, $p_L GF \rightarrow \infty$.

A.2 Proof of Lemmas

Proof of Lemma 1.2.1

Since $c = \sup_{z^* \in [0,1]^{d_Z}} |g_r(z^*)| < \infty$ and g_r is continuous on the compact set $[0,1]^{d_Z}$, we have

$$|g_r(z^*) - g_r(z^0)| \leq 2c \left(\frac{\|z^* - z^0\|}{\delta} \right)^2 + \frac{\varepsilon}{2} \quad \forall z^* \in [0,1]^{d_Z},$$

and thus,

$$\begin{aligned} |B_K(z_1, \dots, z_{d_Z}, g_r) - g_r(z^0)| &= |B_K(z_1, \dots, z_{d_Z}, g_r - g_r(z^0))| && \text{(Binomial Theorem)} \\ &\leq B_K \left(z, 2c \left(\frac{\|z^* - z^0\|}{\delta} \right)^2 + \frac{\varepsilon}{2} \right) \\ &= \frac{2c}{\delta^2} B_K \left(z, \|z^* - z^0\|^2 + \frac{\varepsilon}{2} \right). \end{aligned}$$

It is implied that for $t = t^*$ specifically, we have

$$|B_K(z^0, g_r) - g_r(z^0)| \leq \frac{2c}{\delta^2} \sum_{m=1}^{d_Z} \frac{z_m^0 - (z_m^0)^2}{K_m} + \frac{\varepsilon}{2} \leq \sum_{m=1}^{d_Z} \frac{c}{2K_m \delta^2} + \frac{\varepsilon}{2}$$

since $z_m^0 - (z_m^0)^2 \leq \frac{1}{4}$. Therefore, with equal order for all z_1 to z_{d_Z} , say $\bar{K} := K_1 = \dots = K_{d_Z}$, $\bar{K} \geq (d_Z c) / (\delta^2 \varepsilon)$ implies that $\sup_{z \in \mathbb{R}^{d_Z}} |B_K(z, g_r) - g_r(z)| \leq \varepsilon$. \square

Proof of Lemma 1.2.2

By the Weierstrass approximation theorem, there exist some network structure s under polynomial activation functions such that a continuous underlying function $g(Z)$ can be arbitrarily closely approximated by $g_n^s(Z; W)$. Then under such a polynomial network structure, the convergence $g_n^s(Z_i; \hat{W}) - g_n^s(Z; W) \xrightarrow{p} 0$ follows from the convergence of the trained weights (see, e.g., Allen-Zhu et al., 2019; Zeng et al., 2019), which further implies that $\mathbb{E}[\|g_n^s(Z; \hat{W}) - g(Z)\|_F^2] = o(1)$, given that d_X is finite.

Proof of Lemma A.1

By subtracting and adding terms, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n g_n^s(Z_i; \hat{W}) X_i^\top &= \frac{1}{n} \sum_{i=1}^n g(Z_i) X_i^\top + \\ &\quad \frac{1}{n} \sum_{i=1}^n \{g_n^s(Z_i; \hat{W}) - g(Z_i)\} X_i^\top. \end{aligned} \quad (\text{A.5})$$

Under Assumption 1.2.1, we apply Lemma A.2, letting $A = n^{-1} \sum_{i=1}^n g(Z_i) X_i^\top + (\text{A.5})$, $B = \Sigma_{gX}$, $f(C) = e_i C^{-1} e_j$ for vectors e_i with the i th entry being one and zeros elsewhere, $i, j = 1, \dots, d_X$, and any d_X -by- d_X invertible matrix C with $\min\{a_{ij}, b_{ij}\} < c_{ij} < \max\{a_{ij}, b_{ij}\}$ for all the elements a_{ij} , b_{ij} and c_{ij} of matrices A , B and C , respectively. Then we have $e_i A^{-1} e_j = e_i B^{-1} e_j + \text{tr}(D)$, where $D = \{\partial f(C)/\partial C\}^\top \{(\text{A.5})^* + o_p(1)\}$ for $(\text{A.5})^* = \mathbb{E}[\{g_n^s(Z_i; \hat{W}) - g(Z_i)\} X_i^\top]$. There exists some constant $\mathfrak{c} \in \mathbb{R}$, such that

$$\begin{aligned} \text{tr}(D) &\leq \mathfrak{c} \|D\|_F = \mathfrak{c} \left\| \left\{ \frac{\partial f(C)}{\partial C} \right\}^\top \{(\text{A.5})^* + o_p(1)\} \right\|_F \\ &\leq \mathfrak{c} \left\| \frac{\partial f(C)}{\partial C} \right\|_F \|(\text{A.5})^* + o_p(1)\|_F \\ &\leq \mathfrak{c} \left\| \frac{\partial f(C)}{\partial C} \right\|_F [\|(\text{A.5})^*\|_F + o_p(1)]. \end{aligned}$$

Assumption 1.2.3 and Lemma 1.2.2 together imply that $\|\partial f(C)/(\partial C)\|_F \in \mathcal{O}_p(1)$. Meanwhile, we have $\| [g_n^s(Z_i; \hat{W}) - g(Z_i)] X_i^\top \|_F \leq \|g_n^s(Z_i; \hat{W}) - g(Z_i)\|_F \|X_i\|_F$ by the submultiplicativity of the Frobenius norm. Then applying Lemma 1.2.2, Jensen's inequality, and the convexity of the Frobenius norm yields

$$\|(\text{A.5})^*\|_F \leq \mathbb{E} [\|g_n^s(Z_i; \hat{W}) - g(Z_i)\|_F \|X_i\|_F] = o_p(1).$$

Hence, together with Assumptions 1.2.1 to 1.2.3, Lemma A.1(a) is implied, whence Lemma A.1(b) follows with the fact that d_X is finite and the previous result of $e_i A^{-1} e_j = e_i B^{-1} e_j + \text{tr}(D)$ by Lemma A.2.

Proof of Lemma A.2

First, let $\psi(q) := f(B + q(A - B))$ for $q \in [0, 1]$. Then taking the first order derivative of $\psi(q)$ with respect to q through the matrix argument of the function f yields

$$\begin{aligned}\psi^{(1)}(q) &= \text{tr} \left[\left\{ \frac{\partial f(B + q(A - B))}{\partial (B + q(A - B))} \right\}^\top \left\{ \frac{\partial (B + q(A - B))}{\partial q} \right\} \right] \\ &= \text{tr} \left[\left\{ \frac{\partial f(B + q(A - B))}{\partial (B + q(A - B))} \right\}^\top (A - B) \right].\end{aligned}$$

By the mean-value theorem, there exists some $q \in [0, 1]$, such that

$$\psi(1) - \psi(0) = \psi^{(1)}(q),$$

which is equivalent to

$$f(A) - f(B) = \text{tr} \left\{ \left[\frac{\partial f(C)}{\partial C} \right]^\top (A - B) \right\}.$$

B Appendices of Chapter 2

Notes for Equation 2.2.6

We apply a L_2 least square loss function for the estimation using DNN. We provide the proof why minimizing such loss function leads to achieving the underlying function g . From the law of large numbers (LLN), $\text{Loss}_n^s(Q) := \frac{1}{n} \sum_{i=1}^n \|Y_i - g_n^s(A_i; Q)\|_{L_2}^2 \xrightarrow{P} \mathbb{E}\{[Y_i - \hat{g}_{DNN}(A)]^2\}$, as $n \rightarrow \infty$, where $g_{DNN}(A_i) = g_n^s(A_i; Q)$. Then we have,

$$\begin{aligned} \mathbb{E}\{[Y_i - \hat{g}_{DNN}(A_i)]^2\} &= \mathbb{E}\{(Y_i - \mathbb{E}[Y_i|A_i] + \mathbb{E}[Y_i|A_i] - \hat{g}_{DNN}(A))^2\}, \\ &= \mathbb{E}\{(Y_i - \mathbb{E}[Y_i|A_i] + \hat{k}(A_i))^2\}, \text{ Suppose } \hat{k}(A_i) = \mathbb{E}[Y_i|A_i] - \hat{g}_{DNN}(A_i), \\ &= \mathbb{E}\{(Y_i - \mathbb{E}[Y_i|A_i])^2\} + \mathbb{E}\{(\hat{k}(A_i))^2\} + 2\mathbb{E}\{(Y_i - \mathbb{E}[Y_i|A_i])\hat{k}(A_i)\} \\ &= \mathbb{E}\{(Y_i - \mathbb{E}[Y_i|A_i])^2\} + \mathbb{E}\{(\hat{k}(A))^2\} \\ &\geq \mathbb{E}\{(Y_i - \mathbb{E}[Y_i|A_i])^2\}, \text{ if and only if } \hat{g}_{DNN}(A_i) = \mathbb{E}[Y_i|A_i] = g(A_i). \end{aligned}$$

With an appropriate neural network architecture and the L_2 least square loss function, the estimated function $g_{DNN}(A_i)$ converge to the conditional expectation function $\mathbb{E}[Y_i|A_i]$ or $g(A_i)$ as $n \rightarrow \infty$.

Proof for Lemma 2.2.3

With Equations 2.2.4, we have $g(A) = \beta X + f(\omega)$. Then we have $g[(X_i = 0, \omega_i)] = f(\omega)$. According to Lemma 2.2.2, we have $\mathbb{E}[\|g_n^s[(X_i = 0, \omega_i; \hat{Q})] - g[(X_i = 0, \omega_i)]\|_{L_2}^2] \rightarrow 0$ for all $A_i \in [0, 1]^{d_A}$ as $n \rightarrow \infty$, where $g_n^s[(X_i = 0, \omega_i; \hat{Q})] = f_n^s[\omega_i; \hat{Q}]$ and $g[(X_i = 0, \omega_i)] = f(\omega_i)$. Therefore, $\mathbb{E}[\|f_n^s(\omega_i; \hat{Q}) - f(\omega_i)\|_{L_2}^2] \rightarrow 0$ for all $\omega_i \in [0, 1]^{d_\omega}$ as $n \rightarrow \infty$ holds.

Proof for Theorem 2.2.1

According to Equation 2.2.8, we have

$$\hat{\beta}_{DNN} - \beta = \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \epsilon_i\right) + \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top\right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_i [f_n^s(\omega_i; \hat{Q}) - f(\omega_i)] \right\},$$

With Assumptions 2.2.1 to 2.2.4, we have $n^{-1} \sum_{i=1}^n X_i \epsilon_i \xrightarrow{P} 0$ and $n^{-1} \sum_{i=1}^n X_i X_i^\top \xrightarrow{P} \Sigma_{XX}$ as $n \rightarrow \infty$. With $f_n^s(A_i; \hat{Q})$ satisfying Lemma 2.2.3 and the results from Chen et al. (2020), then $\mathbb{E}[\|X_i [f_n^s(\omega_i; \hat{Q}) - f(\omega_i)]\|_{L_2}^2] \rightarrow 0$ for all $\omega_i \in [0, 1]^{d_\omega}$ as $n \rightarrow \infty$. Then, we have

$\|\hat{\beta}_{\text{DNN}} - \beta\|_{L_2}^2 \xrightarrow{p} \|\mathbb{E}[X_i X_i^\top]^{-1} \mathbb{E}[X_i \epsilon_i]\|_{L_2}^2$ as $n \rightarrow \infty$. Therefore, $\|\hat{\beta}_{\text{DNN}} - \beta\|_{L_2}^2 \xrightarrow{p} 0$ as $n \rightarrow \infty$.

Proof for Theorem 2.2.2

Given Equation 2.2.8, we have

$$\sqrt{n}(\hat{\beta}_{\text{DNN}} - \beta) = \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \epsilon_i\right) + \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i'\right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_i [f_n^s(\omega_i; \hat{Q}) - f(\omega_i)] \right\},$$

Under Assumptions 2.2.1 to 2.2.5, the first part $\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \epsilon_i\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{XX}^{-1} \Omega \Sigma_{XX}^{-1})$ as $n \rightarrow \infty$. The convergence rate of $\hat{f}_n^s(\omega_i; \hat{Q}) - f(\omega_i)$ is shown in Lemma 2.2.1. We have the second part $\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i'\right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_i [\hat{f}_n^s(\omega_i; \hat{Q}) - f(\omega_i)] \right\} \in o_p(n^{-1/2})$, illustrated by Chen et al. (2020). Therefore, we have $\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i'\right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_i [\hat{f}_n^s(\omega_i; \hat{Q}) - f(\omega_i)] \right\} \xrightarrow{p} 0$ as $n \rightarrow \infty$. Then we obtain the results in Theorem 2.2.2.

Notes for Equation 2.2.11 We apply a moment-condition-based loss function for the estimation using DNN. We provide the proof why minimizing such loss function leads to achieving the underlying function g . From the Equation 2.2.11, we have $\frac{1}{n} \sum_{i=1}^n [(Y_i - \hat{\beta} X_i - f_n^s(\omega_i; Q)) \times k(\omega_i)] = \frac{1}{n} \sum_{i=1}^n [(M + \epsilon) \times k(\omega_i) + \frac{1}{n} \sum_{i=1}^n \{g[(X_i, \omega_i)] - g_n^s[(X_i, \omega_i); Q]\} \times k(\omega_i)]$. Given the exogeneity of M and ϵ , we have $\mathbb{E}[M + \epsilon | W] = 0$. Then

$$E[(M + \epsilon)k(\omega)] = E[E[(M + \epsilon)k(\omega) | \omega]] = E[k(\omega)E[(M + \epsilon) | \omega]] = 0.$$

Therefore, from the law of large numbers (LLN), $\frac{1}{n} \sum_{i=1}^n [(M_i + \epsilon_i) \times k(\omega_i)] \xrightarrow{p} 0$ as $n \rightarrow \infty$. Also, we have $\frac{1}{n} \sum_{i=1}^n \{g[(X_i, \omega_i)] - g_n^s[(X_i, \omega_i); Q]\} \times k(\omega_i) = \frac{1}{n} \sum_{i=1}^n \{\beta X_i - \hat{\beta} X_i + f(\omega_i) - f_n^s[\omega_i; Q]\} \times k(\omega_i)$. We prove that when the loss function becomes zero or small enough, a unique estimation for $g_n^s[(X_i, \omega_i); Q]$ will be provided. Suppose there are $\hat{\beta}_1$, $f_{1n}^s(\omega_i)$ and $g_{1n}^s[(X_i, \omega_i); \hat{Q}] := \hat{\beta}_1 X_i + f_{1n}^s(\omega_i; \hat{Q})$. Then $\mathbb{E}[\{g[(X_i, \omega_i)] - g_{1n}^s[(X_i, \omega_i); \hat{Q}]\} \times k(\omega_i)] = \mathbb{E}[(\beta - \hat{\beta}_1) X_i + (f(\omega_i) - f_{1n}^s(\omega_i; \hat{Q})) \times k(\omega_i)]$. Assume $\mathbb{E}[(\beta - \hat{\beta}_1) X_i + (f(\omega_i) - f_{1n}^s(\omega_i; \hat{Q})) \times k(\omega_i)] = 0$, we will have $\hat{\beta}_1 = \beta$ and $f_{1n}^s(\omega_i; \hat{Q}) = f(\omega_i)$ as $k(\omega)$ is an arbitrary function of ω satisfying Assumption 2.2.7. Therefore, $\frac{1}{n} \sum_{i=1}^n \{g[(X_i, \omega_i)] - g_n^s[(X_i, \omega_i); Q]\} \times k(\omega_i) \xrightarrow{p} 0$ as $n \rightarrow \infty$ if and only if $\mathbb{E}[\|g_n^s[(X_i, \omega_i); Q] - g[(X_i, \omega_i)]\|_{L_2}^2] \xrightarrow{p} 0$ in Lemma 2.2.5.

Proof for Theorem 2.2.3

According to Equation 2.2.12, we have

$$\hat{\beta}_{\text{DNN}} - \beta = \frac{1}{n} \sum_{i=1}^n \frac{1}{r} \{g_n^s[(X_i + r, \omega_i); \hat{Q}] - g_n^s[(X_i, \omega_i); \hat{Q}] - r\beta\},$$

With the Lemma 2.2.5, we have $\|g_n^s[(X_i, \omega_i); \hat{Q}] - g[(X_i, \omega_i)]\|_{L_2}^2 \xrightarrow{p} 0$ as $n \rightarrow \infty$. Therefore, $\mathbb{E}[\|g_n^s[(X_i + r, \omega_i); \hat{Q}] - g_n^s[(X_i, \omega_i); \hat{Q}] - r\beta\|_{L_2}^2] \leq \mathbb{E}[\|g_n^s[(X_i + r, \omega_i); \hat{Q}] - g[(X_i + r, \omega_i)]\|_{L_2}^2] + \mathbb{E}[\|g_n^s[(X_i, \omega_i); \hat{Q}] - g[(X_i, \omega_i)]\|_{L_2}^2] \xrightarrow{p} 0$ as $n \rightarrow \infty$. Then we have $\|\hat{\beta}_{\text{DNN}} - \beta\|_{L_2}^2 \xrightarrow{p} 0$ as $n \rightarrow \infty$.